LBL-38594
UC-404

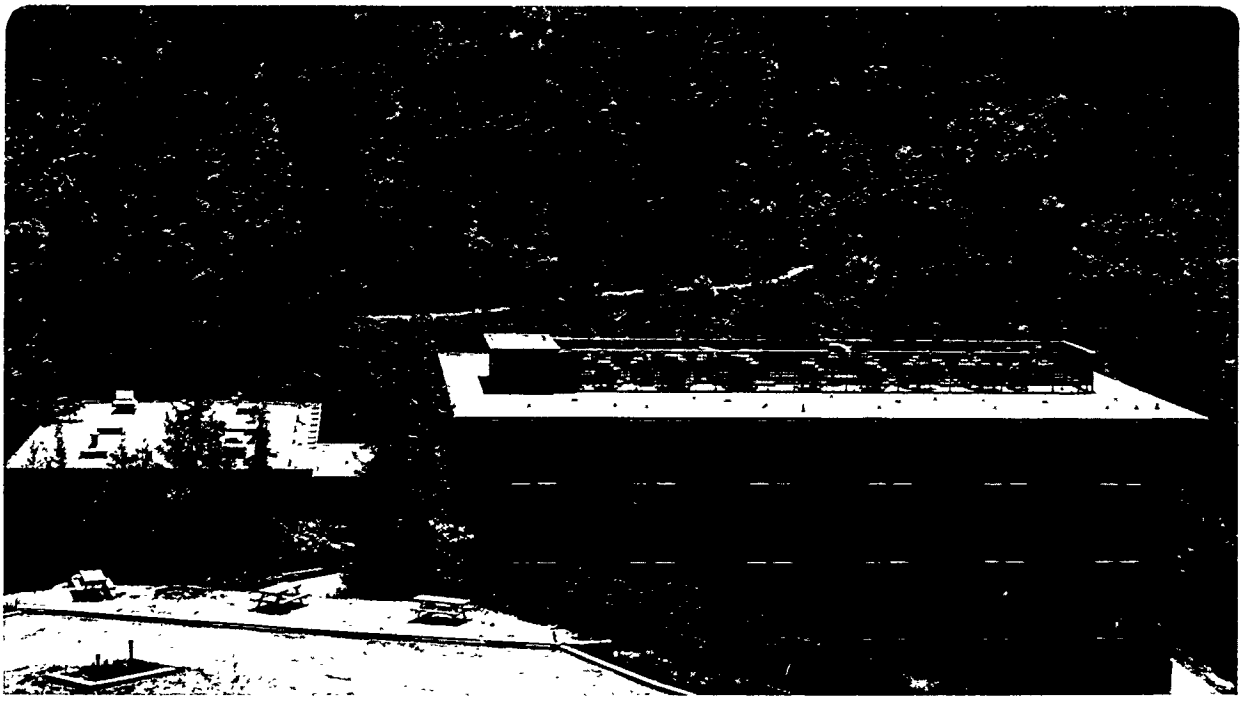
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Materials Sciences Division

Ultrafast Scanning Tunneling Microscopy

D.A. Botkin
(Ph.D. Thesis)

September 1995



Prepared for the U.S. Department of Energy under Contract Number [illegible]

LBL-38594
UC-404

Ultrafast Scanning Tunneling Microscopy

D.A. Botkin
Ph.D. Thesis

Department of Physics
University of California, Berkeley

and

Materials Sciences Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

September 1995

This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098; and by ONR/ARPA under Contract No. N000-14.93.105.36

Ultrafast Scanning Tunneling Microscopy

by David Aaron Botkin

B.A. (University of California, Berkeley) 1989

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Physics
in the
GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA, BERKELEY

Committee in Charge:
Professor Daniel S. Chemla, Chair
Professor Roger Falcone
Professor Jeffrey Bokor

Ultrafast Scanning Tunneling Microscopy

by

David Aaron Botkin

Ultrafast Scanning Tunneling Microscopy

by
David Aaron Botkin
Doctor of Philosophy in Physics
University of California at Berkeley
Professor Daniel S. Chemla, Chair

I have developed an ultrafast scanning tunneling microscope (USTM) based on our novel technique for uniting the stroboscopic methods of ultrafast optics and the technology of scanned probe microscopy (SPM) to obtain nanometer spatial resolution and sub-picosecond temporal resolution. The USTM increases the achievable time resolution of a scanning tunneling microscope (STM) by more than 6 orders of magnitude. This new capability should enable the exploration of mesoscopic and nanometer size systems on time scales which correspond to the period or decay of fundamental excitations.

The USTM consists of a photoconductive switch with subpicosecond response time in series with the tip of a STM. An optical pulse from a modelocked laser activates the switch to create a gate for the tunneling current, while a second laser pulse on the sample initiates a dynamic process which affects the tunneling current. By sending a large sequence of identical pulse pairs and measuring the average tunnel current as a function of the relative time delay between the pulses in each pair, one can map the time evolution of the surface process.

I have used the USTM to measure in the time domain the broadband response of the STM's atomic size tunnel barrier to frequencies from tens to hundreds of gigahertz. The amplitude of the USTM signal amplitude decays linearly with the conductance of the tunnel junction; hence the spatial resolution of the time-resolved signal is comparable to the resolution of a conventional STM. The geometrical capacitance of the junction does not appear to play an important role in the measurement. However, a capacitive effect intimitely related to the process of tunneling contributes significantly to the measured signals. This effect may define or limit the ultimate resolution of the USTM.

CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Many people helped me during the course of my dissertation research. Frank Ogletree and Miquel Salmeron taught me how to design and use an STM and were instrumental in getting the project started. Jeff Beeman helped to make our first STM tips and taught me how to manipulate very small objects. Tony Levi taught me microlithography in a whirlwind week at Bell Labs. Chandu Karadi taught me how to make photoconductive switches and was the source for many important insights into the physics of the experiment. Taekjip Ha helped with innumerable problems in the lab. Jennifer Glass has helped with recent improvements to the microscope. Stephan Glutch always knew (and told me) the answers to my math questions.

I would like to thank my advisor, Daniel Chemla for his support and encouragement. He taught me a great deal of physics and physical intuition during our 5 years together. I would especially like to thank Shimon Weiss, who acted like a second advisor to me. Our almost daily discussions and interaction in the lab taught me how to be an experimental physicist. Shimon also invented the original concept for the experiment and worked extensively with me in the lab during our early measurements.

I am grateful to the other residents of Building 2 (there are too many to mention individually) who provided animated lunchtime discussions on every topic imaginable and who made the environment lively and creative. Thank you also to my friends on the Cal Racquetball Team and to Chris Kutz for getting me out of the lab and on to a racquetball or squash court once in a while.

My wife, Lee Anna, gave me the inspiration and encouragement necessary to survive graduate school. I would also like to thank her for her help in reviewing this manuscript.

Finally, I would like to thank my parents, who fostered my interest in physics (and let me shine lasers on the neighbors windows). They have provided constant support and encouragement for as long as I can remember.

This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098; and by ONR/ARPA under Contract No. N000-14.93.105.36.

"Through purely logical thinking we can attain no knowledge whatsoever of the empirical world."

Albert Einstein

Chapter 1
INTRODUCTION

1.1. The need for time and space resolutions

Solid surfaces exhibit an extraordinary range of dynamic behavior. Electrons are in continuous motion at the Fermi velocity, from 1-10 Å/fs, and diffuse or move in response to applied fields at somewhat lower speeds. Photons are absorbed and emitted, launching plasmons and phonons. Surface and overlayer phonons propagate at velocities ranging between 0.001-0.1 Å/fs[1]. Molecules land, react with other molecules, diffuse, are adsorbed, and desorb. Adsorbed molecules vibrate in place with periods on the order of 100 femtoseconds and with vibrational decay times on the order of a few picoseconds[2]. These processes affect the microscopic and macroscopic properties of matter. To study them fully in a field of view substantially smaller than a micron requires a new class of imaging tools with simultaneous high spatial and high temporal resolution.

High spatial resolution is particularly important for investigations of dynamics in the novel mesoscopic systems which may play an important role in future electronics and optoelectronics applications. For example, carrier transport in mesoscopic semiconductor structures and electric field and voltage wave front propagation at metal-semiconductor or p-n interfaces exhibit extreme variation over length scales much smaller than 1 μm. In fact, the physics of the transient processes governing these structures is only known for values of observables averaged over length scales on the order of a micron. The physics is poorly understood on an atomic length scale, especially in the subpicosecond temporal regime. Techniques which can provide high spatial and temporal resolution also are becoming increasingly important, particularly as new design rules for IC manufacture bring smaller signal carrying lines and higher clock frequencies[3].

Until we proposed and demonstrated the concept of ultrafast scanning probe microscopy (USPM)[4,5], which combines the high spatial resolution of scanning probe microscopy (SPM) with the high temporal resolution of ultrafast optical techniques, no general tools existed which could attain simultaneous picosecond time resolution and nanometer spatial resolution. Existing techniques, like voltage contrast scanning electron microscopy[6] and photoemission sampling[7], could achieve several picosecond time resolution and on the order of one hundred nanometer spatial resolution. A recent and novel technique, the femtosecond field emission camera, allowed the continuous observation in real time of a molecule adsorbed at the end of a field emission tip with picosecond and sub-nanometer resolution[8], but it had little practical application.

Nanometer spatial resolution and sensitivity to a very large class of surface dynamics is readily available with the techniques of SPM. However, while SPM can provide up to atomic spatial resolution, its temporal resolution is rather limited. Conversely, optical techniques like electro-optic sampling[9] and time resolved absorption spectroscopy[10], which can resolve dynamics on a femtosecond time scale, have spatial resolutions limited to about 500 nm.

1.2. Scanned Probe Microscopy

The invention of the scanning tunneling microscope (STM)[11] revolutionized surface science by enabling the direct observation of nanometer scale features on surfaces, and it catalyzed the development of a host of other scanned probe techniques which have been used to produce images of various surface properties with up to atomic resolution. These techniques include the atomic force microscope (AFM)[12] ,the electrostatic force microscope (EFM)[13], the magnetic force microscope (MFM)[14], and the scanning near field optical microscope (SNOM or NSOM)[15]. Each of the variations of SPM relies on a highly local interaction between a small probe and the sample surface to achieve its high resolution. A typical image records the strength of this interaction as a function of the lateral position of the tip relative to the sample.

The interaction is always nonlinear with respect to variation of spatial parameters (this is what it means to be spatially local), and it is often nonlinear with respect to other parameters as well. In STM, for example, the measured property is the current of electrons which tunnels across the gap between the tip and the sample. This tunneling current varies exponentially with changes in tip to sample separation and varies like $aV_{bias} + (bV_{bias})^3$ with changes in the bias voltage across the gap. In the EFM, the force on the cantilever is proportional to the square of the voltage between tip and sample and varies inversely with the distance between tip and sample.

It is possible to extend the capabilities of SPM to the time domain by increasing the bandwidth of the measurement. The first such attempt was by Binnig and co-workers, who noticed that noise spikes in the tunnel current could arise from diffusion of atoms under the STM tip[16]. At around the same time, Quate developed an STM which could produce video rate images on atomically flat graphite, primarily as a means to reduce the influence of low frequency and 1/f noise in the measurement[17]. More recently, several groups have studied diffusion at

millisecond time scales[18] and surface reconstruction on a time scale of seconds[19]. Mamin and colleagues developed cantilevers with extremely high resonance frequencies to achieve 1.2 MBit/s reading rates in data storage applications with AFM[20]. The same group also pushed the electronic and mechanical limits of STM to obtain scanning rates of 1 mm/s (with nanometer resolution) and video rate imaging on atomically rough surfaces[21].

Mechanical and electronic constraints limit the time resolution of conventional SPM to on the order of milliseconds for images and on the order of microseconds for measurements at a single point. In scanning force microscopy the ultimate limit is set by the resonance frequency of the cantilever, which is at best a few MHz[20]. In STM, the intrinsic response time of the junction is thought to be limited by tunneling time to the order of femtoseconds[22], but the bandwidth of the preamplifier required to amplify the signal to measurable levels is usually below 100 kHz and is at best a few MHz. Real-time measurements in SPM thus have time resolution on the order of microseconds or slower.

1.3. Ultrafast Optics

Researchers in the field of ultrafast optics long have evaded the same types of electrical and mechanical constraints through the use of stroboscopic techniques. In a stroboscopic measurement, a particular temporal slice of a fast process is singled out and then its average amplitude is measured with slow detectors and electronics. The fast process is generated repetitively, and the signal is recorded as a function of the position of the temporal window which selects the measured slice. The position of the window can be scanned as slowly as is necessary to build up a measurable signal; therefore the time resolution is determined by the width of each temporal slice rather than by the speed of the measurement electronics. The general concept dates back at least as far as microsecond resolution spark photography in the 1850s[23]. With modern modelocked lasers, picosecond resolution is routine, and resolution of 6 fs, very near the limit set by the uncertainty principle for optical frequencies, has been demonstrated[24].

The quintessential technique of ultrafast optics is the pump-probe measurement, in which a pump laser pulse initiates a transient process in a medium, and a probe pulse samples the process after a specified time delay. The measurement can be viewed as a correlation mediated by a nonlinear interaction between the pump induced transient and the probe pulse. It is the nonlinearity which mixes high frequencies in the transient down to quasi-DC levels that we can measure.

1.4. Ultrafast Scanning Probe Microscopy

1.4.1. Basic Concept

The nonlinear interaction between tip and sample in SPM provides the means to marry the techniques of ultrafast optics with those of SPM[4,5]. The basic concept is illustrated in Figure 1.1. A pump laser pulse generates the transient effect one wishes to observe. The signal which would normally result from this transient is too fast to measure within the bandwidth of the SPM. However, the transient and a probe laser pulse induced tip response are mixed by the nonlinear interaction to produce quasi-DC terms which can easily be measured. The spatial resolution should be the same or comparable to the usual resolution of the SPM, since the mixing occurs at the spatially local nonlinearity.

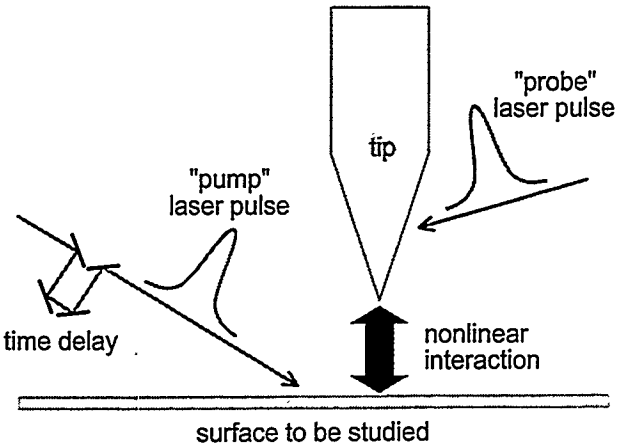


Figure 1.1. The basic concept behind USPM - correlation via nonlinear interaction. The two laser pulses modulate the sample and the tip responses respectively. The nonlinear interaction produces a signal dependent on temporal separation between the laser pulses.

An intuitive picture of the mechanism is this: the probe pulse and the nonlinearity combine to form a short gate for the transient process. The width of the gate, and therefore the time resolution of the measurement, depends on the temporal width of the probe induced response and the time response of the nonlinearity. For example, in STM, a probe beam focused on the tunneling gap can be designed to produce a photothermal modulation of the gap width or to modulate the voltage bias across the gap. In the latter case, the response time is the inverse plasma frequency, so the time resolution will be on the order of the (much longer) pulse width. In the former case, the probe induced response will have a temporal width on the order of microseconds[25] and will determine the time resolution of the measurement.

The nonlinearity which provides the time resolution does not have to be inherent in the tip-sample interaction. For example, we can introduce an artificial nonlinearity such as a photoconductive switch which gates the signal

after it has been generated by the intrinsic tip sample interaction. In this case, the spatial resolution is not affected, but we now have the capability to study the dynamic response of the intrinsic nonlinearity as well as dynamic processes on the sample.

The USPM technique in principle is not limited to measurements at single points. While the time delay between pump and probe is fixed, the tip can scan the surface and produce an image, creating a snapshot of an ultrafast process. A series of such snapshots created for several time delays and viewed in rapid succession form a movie of surface dynamics.

1.4.2. Previous and ongoing work in the field

There have been several successful attempts to link stroboscopic techniques with SPM. The first attempt to combine an STM with ultrafast lasers, by Hamers and Cahill, relied on the nonlinear interaction of two laser pulses on the sample and used the STM as a slow detector[26]. The technique demonstrated nanosecond time resolution but was not a general method for introducing time resolution to STM.

Independently and in parallel with our own work, Nunes and Freeman extended the experiment of Hamers and Cahill by mixing 100 ps voltage pulses with the STM's nonlinear I-V curve to demonstrate 100 ps resolution. By capacitively coupling a probe voltage directly to the tip, they demonstrated 1 ns resolution in a sample-independent configuration[27]. Nunes and Freeman also used the exponential I-Z characteristic of the STM via gap distance modulation to obtain on the order of 10 nanosecond resolution in the measurement of an acoustic pulse[28].

Kim and coworkers demonstrated 2 picosecond resolution in electrical sampling measurements using a freely positionable contact probe in series with a photoconductive switch[29]. Their technique obtains spatial resolution on the order of 0.1 μm, determined by the size of the contact area. Takeuchi and Kasahara used a technique nearly identical to ours to demonstrate 300 ps resolution in STM after we demonstrated 2 ps resolution[30].

During the process of editing this thesis, Groeneveld and co-workers demonstrated several picosecond resolution by following our technique with a simple and elegant STM tip design[31]. This tip design is similar to one we tried and abandoned as unsuitable for spatial imaging, and I believe the new design remains unsuitable for spatial imaging. However, the work is very promising and problems with the tips probably can be overcome. Significantly, Groenveld's recent results have now corroborated our measurements and observations over the past few years.

Some of the best work with USPM has been performed by the Bloom group at Stanford in parallel with our work. Using an all electronic sampling technique, they have used the EFM as a square law detector to obtain 5 ps time resolution[32]. At the time of our first results, they had demonstrated nonlinear mixing of two signals propagating on a sample with a resolution of 100 ps[33]. The spatial resolution of their technique is on the order of 0.1 μm.

1.5. Ultrafast Scanning Tunneling Microscopy

1.5.1. Why STM?

Although the basic technique illustrated in Figure 1.1 can be applied to a variety of SPM techniques, STM has several advantages which make it very desirable for high time resolution measurements. STM is sensitive to the largest variety of surface dynamics of any of the SPM techniques. It is sensitive to electronic changes at the Fermi level, which can be caused by changes in local voltage or charge density, and it is sensitive to sub-angstrom variations in surface height, which can be caused by surface topography or phonons. The STM also has the highest lateral spatial resolution of any scanned probe technique. In addition to the versatility and resolution of STM, the physics of the high frequency response of the STM tunnel junction is interesting in its own right. An ultrafast STM offers the potential to study high frequency transport through the atomic sized channel formed at the STM's tunnel junction and offers the remote but tantalizing possibility of addressing the hotly debated issue of tunneling time.

The STM has an intrinsically fast response time. Three fundamental effects limit the ultimate time resolution of measurements made with an STM. These effects are the tunneling time, the inverse frequency of the plasmon modes in the tip and sample, and the *RC* time constant of the junction. *R*, the spreading resistance in series with the junction, has a value on the order of a few ohms[34]. A conservatively large value for the capacitance is a few fF, so the *RC* time constant is on the order of 10 fs. A reasonable estimate for the tunneling time across a 5 Å gap (typical for an STM) is

$$\tau_{tunneling} = \frac{d}{v_{Fermi}} \approx \frac{5\,\text{Å}}{1\,\text{Å}/\,\text{fs}} = 5\,\text{fs}.$$

The plasma frequency in metals is in the ultraviolet, so the inverse plasma frequency is on the order of 1 fs. The ultimate time resolution achievable in STM is therefore on the order of 10 fs.

1.5.2. Possible Approaches

As discussed previously, the only way to realize such resolution with existing electronics is through a pump-probe technique which mixes the immeasurable high frequencies down to quasi-DC levels. In a pump-probe experiment, one measures the time-averaged signal,

$$S(\tau) = \frac{1}{2T_0} \int_{-T_0}^{T_0} I(t,\tau)dt ,$$

where $I(t, \tau)$ is a transient signal which results from the combined action of pump and probe excitations, τ is the relative time delay between the pump and probe, and $2T_0$ is the interval over which the signal is measured. The transient signal $I(t,\tau)$ must depend nonlinearly on the pump and the probe. Otherwise, we can independently sum the responses from each, and the time separation between the two responses becomes irrelevant. Usually the experiment is performed repetitively, and the system relaxes to its initial state between sets of pump and probe excitations. In such a case, $I(t,\tau)$ is periodic with the repetition rate of the measurement, so $2T_0$ can be taken to be a single repetition period.

In STM, the relevant transient signal is the tunneling current, and the measured signal can be either the time-averaged tunneling current or the reaction of the STM feedback loop as it compensates changes in the time-averaged current. The element which nonlinearly mixes signals from the pump and the probe can be a native nonlinearity in the STM junction or can be a foreign nonlinearity introduced in the junction or elsewhere.

The most conceptually simple approach to time resolved measurements with STM is to use one of the natural nonlinearities of the STM junction, the *I-V* and *I-Z* responses, to mix signals generated by the pump and probe to obtain time resolution. As discussed in the previous section, some demonstrations of the feasibility of these techniques have been performed with time resolution in the nanosecond regime. Although in theory these methods can achieve sub-picosecond resolution, in practice it is extremely difficult to advance much beyond 100 ps resolution due to limitations in signal strength, noise performance, and the attainable speed of modulation of the nonlinearity. Appendix 1 discusses the theory and limitations of these techniques in more detail.

A more practical approach leads to the instrument we have developed, the Ultrafast Scanning Tunneling Microscope (USTM)[5]. Figure 1.2 schematically illustrates the method. The primary ingredient in the USTM is an artificial nonlinearity, a fast photoconductive switch fabricated in series with the STM junction. The switch functions as a gate for the tunneling current which opens and closes with picosecond time resolution. This method permits us to study transient processes on surfaces as well as the dynamic response of the tunnel junction itself. More importantly, it is possible to implement our method with picosecond time resolution and reasonable signal to noise.

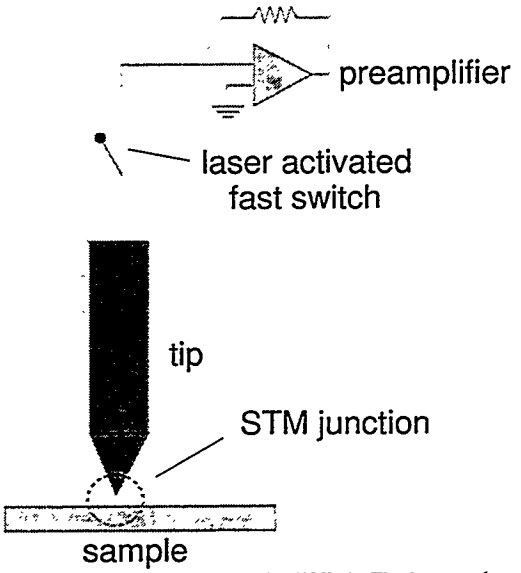


Figure 1.2. A schematic of the USTM. The laser-activated (photoconductive) switch creates a fast gate for the tunneling current.

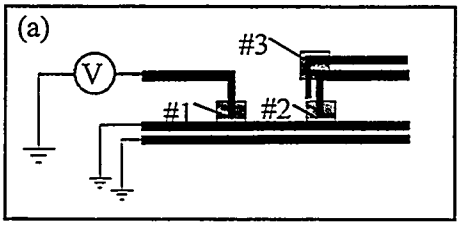
1.5.3. Proof of Principle

Because the amplitude of the signal expected in a USTM measurement is difficult to predict, we initially performed a simulation of the measurement of a picosecond voltage pulse. The simulation was performed in a planar geometry using a silicon-on-sapphire (SOS) pad to model the impedance of the tunnel junction. The experimental arrangement is shown in Figure 1.3a. The gray squares represent fast photoconductive switches, and the black lines represent metal electrodes on a sapphire substrate.

A pump laser pulse at switch #1 launched a voltage pulse on a coplanar transmission line, and a probe laser pulse at switch #3 interrogated the current which "tunneled" through switch #2. The ohmic resistance of the 15 μm silicon gap which forms switch #2 (several 10's of MΩs) was comparable to that of an STM gap in normal operation. The distance between switch #2 and switch #3 was 200 μm.

The shape of the pulse on the sample was measured via conventional photoconductive sampling at switch #2. This technique will be discussed in more detail in the chapters which follow. Figure 1.3b compares the "tunneling" signal, indicated by a solid line, to the conventionally measured signal, indicated by a dotted line. The amplitude of

the "tunneling" signal is multiplied by a factor on the order of 10^3, and the time axis has been shifted so that the peaks of each signal overlap. The average current at the peak of the "tunneling" signal is slightly less than 10 pA. The signal has a large AC component, which indicates capacitive coupling. In fact, this signal is qualitatively similar to USTM measurements performed while the STM tip is out of tunneling range.



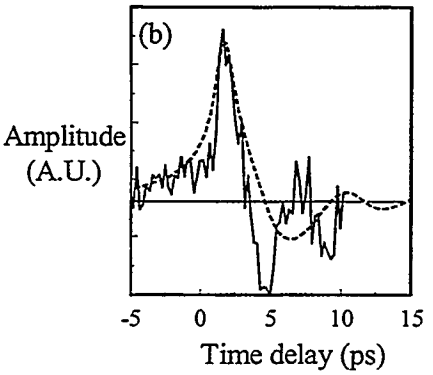


Figure 1.3. Proof of principle USTM simulation experiment on SOS. (a) sample layout and biasing geometry. (b) Dotted line: conventional photoconductive sampling between points #1 and #2. Solid line: photoconductive sampling between points #1 and #3.

1.6. Outline of Thesis

This thesis concerns the development and successful demonstration of the USTM. Because the work is primarily experimental, it is appropriate that Chapter 2 discusses the design and construction of the instrument. I describe the two STM heads we have constructed, the optical system, time delay scanners, and software used for the measurement. I detail our progress in developing suitable photoconductively switched tips, and discuss the methods we currently use.

In the third chapter, I present the results of single point, time resolved tunneling measurements of voltage pulses propagating on a coplanar transmission line. I discuss some of the essential features of the time resolved traces, including their variation with tip-sample separation, voltage, and tip length. From our observations, I show that the time resolution of our USTM technique is less than one picosecond and that the spatial resolution is better than 100 Å. Substantial care is required to identify and reject spurious signals, so the third chapter also discusses the precautions we take to ensure that the measured signal arises from tunneling rather than from geometrical capacitance or antenna effects.

The fourth chapter discusses the applicability of circuit models to describe USTM measurements. Although a circuit model is only an approximate solution to the complicated electrodynamic boundary conditions imposed by the USTM tip, simple models can predict many of our empirical results. Using these models as a guide, I discuss the response of the STM tunnel gap to high frequencies. The tunnel junction behaves as if it has an anomalously small capacitance which is intimately related to the process of tunneling rather than to the conventional geometrical capacitance. The final part of the chapter discusses possible explanations for this surprising result as well as related work in the literature.

The fifth and final chapter describes some of the experiments now in progress with the USTM. These include an attempt to create picosecond scale movies and a measurement of the change in response of the tunnel gap as it collapses to an ohmic contact. The chapter discusses the improvements necessary to make the USTM a versatile laboratory tool, and outlines some of the exciting problems that the USTM may be used to address in the future.

Chapter 1 References:

1. W. Kress, F. W. de Wette (Eds.), *Surface phonons*, Springer Series in Surface Science, Vol. 21, Ch. 8, Springer-Verlag, Berlin/Heidelberg (1991).
2. Andrew Zangwill, *Physics at Surfaces*, Ch 14, Cambridge University Press, Cambridge (1988).
3. C. Reeves, "The uses of scanning electron microscopy for studying semiconductor devices", *Intl. J. Elect.*, 77, p. 919 (1994).
4. David Botkin, Daniel S. Chemla, D. Frank Ogletree, and Shimon Weiss, "Ultrafast Scanning Probe Microscopy", US Patent No. 5,416,327, May 16, 1995.
5. S. Weiss, D. F. Ogletree, D. Botkin, M. Salmeron, and D. S. Chemla, "Ultrafast Scanning Tunneling Microscopy", *Appl. Phys. Lett.*, 63, p. 2567 (1993).
6. J. Bokor, A. M. Johnson, R. H. Storz, and W. M. Simpson, "High-speed circuit measurements using photoemission sampling", *Appl. Phys. Lett.*, 49, p. 226 (1986) and R. B. Marcus, A. M. Weiner, J. H. Abeles, and R. S. D Lin, "High-speed electrical sampling by fs photoemission", *Appl. Phys. Lett.*, 49, p. 357 (1986).
7. See, for example, D. Winkler, R. Schmitt; M. Brunner, and B. Lischke, "Flexible picosecond probing of integrated circuits with chopped electron beams", *IBM J. Res. Develop.*, 34, p. 189 (1990).
8. Harry Heinzelmann, Fumiya Watanabe, Gary M. McClelland, "Observing the motion of a single adsorbed atom with picosecond and subnanometer resolution", *Phys. Rev. Lett.*, 70, p. 3611 (1993)
9. J. A. Valdmanis, G. A. Mourou, and C. W. Gabel, "Picosecond electro-optic sampling system,", *Appl. Phys. Lett.*, 41, p. 211 (1982).
10. W. Kaiser (Ed.), *Ultrashort Laser Pulses and Applications*, Ch. 4, Topics Appl. Phys. Vol. 60, Springer-Verlag, Berlin/Heidelberg (1988).
11. G. Binnig and H. Rohrer, "Scanning Tunnelling Microscopy", *Helvetica Physica Acta*, 55, p. 726 (1982).

12. G. Binnig, C. F. Quate, and C. Gerber, "Atomic force microscope", *Phys. Rev. Lett.*, **56**, p. 930 (1986).
13. Y. Martin, D. W. Abraham; and H. K. Wickramasinghe, "High-resolution capacitance measurement and potentiometry by force microscopy", *Appl. Phys. Lett.*, **52**, p. 1103 (1988).
14. Y. Martin, and H. K. Wickramasinghe, "Magnetic imaging by 'force microscopy' with 1000 Å resolution", *Appl. Phys. Lett.*, **50**, p.1455 (1987).
15. D. W. Pohl, W. Denk, and M. Lanz, "Optical stethoscopy: Image recording with resolution lambda /20", *Appl. Phys. Lett.*, **44**, p. 651 (1984).
16. G. Binnig, H. Fuchs, E. Stoll, "Surface diffusion of oxygen atoms individually observed by STM", *Surf. Sci. Lett.*, **169**, p. L295, (1986).
17. A. Bryant, D. P. E. Smith, and C. F. Quate, "Imaging in real time with the tunneling microscope", *Appl. Phys. Lett.*, **48**, p 832 (1986).
18. see, for example, J.C. Dunphy; P. Sautet, D. F Ogletree, O. Dabbousi, and M.B. Salmeron, "Scanning-tunneling-microscopy study of the surface diffusion of sulfur on Re(0001)", *Phys. Rev. B*, **47**, p. 2320, (1993).
19. see, for example, L. Eierdal, F. Besenbacher; E. Laegsgaard, and I. Stensgaard, "Interaction of oxygen with Ni(110) studied by scanning tunneling microscopy", *Surface Science*, **312**, p. 31 (1994).
20. H.J. Mamin, B.D. Terris, L.S. Fan, S. Hoen, R.C. Barrett, and D. Rugar, "High density data storage using proximal probe techniques", unpublished
21. H. J Mamin, H. Birk; P. Wimmer, and D. Rugar, "High-speed scanning tunneling microscopy: principles and applications", *J. Appl. Phys.*.**75**. p. 161 (1994).
22. M. Buttiker and R. Landauer, *IBM J. Res. Develop.* **30**, p. 451 (1986) and A.A. Lucas, P.H. Cutler, T. E. Feuchtwang, T.T. Tsong, T. E. Sullivan, Y. Yuk. H. Nguyen, and P. J. Silverman, *J. Vac. Sci. Technol. A*, **6**, p. 461, (1988).
23. S. L. Shapiro (ed.) *Ultrashort Light Pulses: Picosecond Techniques and Applications*, Topics in Applied Physics, Vol **18**, Ch. 1, Springer-Verlag, Berlin/Heidelberg (1988).
24. H. L. Fragnito, J.-Y. Bigot, P. C. Becker, and C. V. Shank, "Evolution of the vibronic absorption spectrum in a molecule following impulsive excitation with a 6 fs optical pulse", *Chem. Phys. Lett.*, **160**, p. 101-4 (1989).
25. Nabil M. Amer, Andrew Skumanich, and Dean Ripple, "Photothermal modulation of the gap distance in scanning tunneling microscopy", *Appl. Phys. Lett.*, **49**, p. 137 (1986).
26. R.J Hamers and David G. Cahill, "Ultrafast time resolution in scanned probe microscopies", *Appl. Phys. Lett.* **57** (19), p 2031 (1990).
27. G. Nunes and M. R. Freeman, "Picosecond resolution in scanning tunneling microscopy", *Science*, **262**, p. 1029 (1993).
28. M.R. Freeman and G. Nunes, "Time-resolved scanning tunneling microscopy through tunnel distance modulation", *Appl. Phys. Lett.* **63**, p. 2633 (1993).
29. J. Kim, S. Williamson, J. Nees, S. Wakana, and J. Whitaker, "Photoconductive sampling probe with 2.3-ps temporal resolution and 4-mV sensitivity", *Appl. Phys. Lett.*, **62**, p. 2268 (1993).
30. Koichiro Takeuchi and Yukio Kasahara, "High Speed optical sampling measurement of electrical waveform using a scanning tunneling microscope", *Appl. Phys. Lett.* **63**, p. 3548, (1993).
31. R. H. M. Groeneveld, Th. Rasing, L. M. F. Kaufmann, E. Smalbrugge, J. H. Wolter, M. R. Melloch, and H. van Kempen, "New optoelectronic tip design for ultrafast scanning tunneling microscopy", to be published in *J. Vac. Sci. Technol. A.*
32. A.S. Hou, F. Ho and D.M. Bloom, "High-speed integrated circuit probing using a scanning force microscope sampler", *Electron. Lett.*, **30**, p. 560 (1994).
33. F. Ho, A.S. Hou, and D.M. Bloom, "Picosecond electrical sampling using a scanning force microscope", *Electron. Lett.*, **28**, p. 2302 (1992).
34. Mordehai Heiblum, Shihyuan Wang, John R. Whinnery, and T Kenneth Gustafson, "Characteristics of Integrated MOM Junctions at DC and at Optical Frequencies", *IEEE J. Quant. Elec.*, **QE-14**, p. 159 (1978).

Chapter 2
INSTRUMENT DESIGN

2.1. Introduction

The USTM combines two complicated experimental systems: an STM head with its associated electronics, and a short pulse laser system with its associated optics, time-delay generator, and electronics. The link between the two and the linchpin of the whole system is the tip assembly. The tip assembly contains the STM tip and the photoconductive switch that provides temporal resolution.

There are other, unintended links. The most serious problem is mechanical coupling from the time delay stage to the STM junction. This produces noise synchronized with and therefore indistinguishable from the true signal. Other sources of mechanical noise, such as the cooling system for our laser, reduce the stability of the STM and generally reduce signal to noise in our measurements. The presence of these noise sources constrain the design of the STM head.

The addition of lasers and the necessity of aligning them further constrains the design of the STM. The tip and sample now must be optically accessible and the lateral position of the tip relative to the sample must be controllable on a micron scale. This chapter describes the components of the USTM and discusses the motivation behind and the process involved in the design and creation of the components.

2.2. Tip Assembly

The construction of photoconductively switched tips is possibly the most difficult and critical part of the whole endeavor and was also our starting point. The goal of picosecond time resolution restricts us to low temperature grown GaAs (LT-GaAs)[1] or Silicon-on-Sapphire (SOS)[2] technologies for the switches. Picosecond time resolution also requires that the distance between tip and switch be very short to avoid dispersion. SOS switch technology was chosen because it conveniently allows us to illuminate the switch through the transparent sapphire substrate while the switch is held very close to the sample.

2.2.1. Photoconductive switches

A typical switch consists of a 60 x 30 μm rectangle of Si on a sapphire substrate which bridges the gap between two metal contacts. The undoped SOS wafers that we use, 0.6 μm Si on 0.5 millimeter sapphire substrates, are available commercially from Union Carbide. The switches are defined by standard lithographic techniques and are prepared in an array which can be diced into individual chips, each containing one switch very close to the chip's edge.

In the first step, the Si squares which will form switches are protected by photoresist and the remaining Si is etched away. Removal of the excess Si dramatically reduces dark current. Metal contacts are defined with a liftoff mask. The metal lines are usually Cr (300 Å) /Au (3000 Å) and vary from 20 to 60 μm wide. The gap width varies between 3 and 15 μm. In practice, the temporal response of a switch varies little with geometry within the parameters mentioned above. Different metallizations, like Ti (300 Å) /Au (3000 Å) or Al (3000-5000 Å) also are equally effective.

We follow the recipe of Grischkowsky and implant the Si with $10^{15}/cm^2$ O^- at 200 keV and then $10^{15}/cm^2$ O^- at 100 keV to reduce the carrier lifetime to below one picosecond[3]. The implantation introduces a high density of defects and traps to the first few hundred angstroms below the surface of the material. It also is thought to make the contacts ohmic, which may lead to a more uniform carrier distribution in the switch and ensures that an applied voltage is dropped across the switch instead of across Schottky barriers at the contacts. Cooling the substrate with chilled water during implantation, which makes the damaged region more amorphous, seems to produce faster switches, but cooling with liquid nitrogen seems to provide no additional advantage and costs quite a bit more. In principle, the choice of ion implanter should make no difference, but the original one that we used at LBL consistently produced non-functioning switches. As a result, most of our ion implantation was done at IICO in Santa Clara.

The "off" resistance of a switch is typically on the order of 30 to 100 MΩ, depending on the width of the gap. Anomalously low "off" resistances of less than 10 MΩ usually indicate that the switch was not properly implanted. (Improperly implanted or unimplanted switches also may develop photovoltage at the metal contacts and have average "on" resistances which are very insensitive to laser position.) With a 25 mW, 800 nm laser beam focused to a diameter of 3-5 μm on the switch, the average ("on") resistance of the switch becomes on the order of 1 - 5 MΩ. The peak "on" resistance is easily estimated from the following formula for the switch conductances:

$$G_{on} * T_{on} + G_{off} * T_{off} = G_{avg} * (T_{on} + T_{off}).$$

In this formula, T_{on} represents the amount of time that the switch is on during the period of measurement, and T_{off} is

the amount of time the switch is off. G_{on} and G_{off} are the conductances in the "on" and "off" states. Since the laser pulse arrives at a repetition rate of 76 MHz, T_{off} is approximately 13 ns. T_{on} is given approximately by the carrier lifetime of one picosecond, so the resistance at the peak of the laser pulse is between 75 and 400 Ω.

The capacitance across the switch can be estimated from simple models which approximate the electrode geometry. Modeling the electrodes as parallel plates separated by half air and half Si, we estimate the geometrical capacitance to be

$$C_{switch} \approx \frac{\varepsilon A}{d} \quad F = \left(8.85 \cdot 10^{-12} F/m\right) \cdot \left(\frac{12}{2}\right) \cdot \frac{\left(0.3 \cdot 10^{-6} m\right)\left(40 \cdot 10^{-6} m\right)}{5 \cdot 10^{-6} m} \approx 2 \cdot 10^{-16} F .$$

This model probably underestimates the capacitance by neglecting the effects of fringing fields. However, the effective dielectric constant, which increases the capacitance, may be overestimated since most of the curvature in the electrodes occurs in the air gap. An alternate model of the electrodes as parallel cylindrical electrodes with radii r equal to the thickness of the metal lines and length l equal to the width of the lines yields

$$C_{switch} \approx \frac{2\pi\varepsilon}{\cosh^{-1}\left(\frac{d^2-2r^2}{2r^2}\right)} \cdot l \quad F = 2\pi \cdot \left(8.85 \cdot 10^{-12} F/m\right) \cdot \left(\frac{12}{2}\right) \cdot \frac{40 \cdot 10^{-6} m}{\cosh^{-1}\left(\frac{\left(5 \cdot 10^{-6} m\right)^2}{2\left(0.15 \cdot 10^{6} m\right)^2}\right)} \approx 2 \cdot 10^{-15} F \text{ Hence,}$$

a reasonable estimate for the static capacitance of the switch is on the order of 1 fF.

The magnitude of the photoconductive response, and therefore the sensitivity of the switch, depends on the laser intensity and depends crucially on the laser spot position on the switch. A 1000 Å lateral motion of the laser beam on the photoconductive switch can decrease photoconductivity by a few percent.

2.2.2. Tips

In order to get stable tunneling with nearly atomic resolution in air, a STM's tip needs to be mechanically rigid, chemically clean, and relatively sharp. In a standard STM, the tip is at least several millimeters long and can be freshly prepared before each use. Tips are often electrochemically etched to produce a very small radius of curvature, on the order of tens of nanometers. However, tips made of hard metals also can be prepared simply by cutting with a blunt pair of shears. The cutting fractures the end of the tip, leaving a very sharp region suitable for tunneling. The best tip materials for STMs operating in air are Pt-Rh and Pt-Ir alloys, which do not oxidize and are fairly hard. Materials which oxidize, such as tungsten and aluminum, should usually be avoided. Soft materials such as gold and pure platinum tend to deform during tunneling and also may degrade quickly with time.

In the USTM, we have the unique requirement that the tip be much shorter than one millimeter. There are at least four reasons for this. First, a metal wire without a ground plane is lossy and dispersive for high frequency pulse propagation. A one picosecond Gaussian pulse (which has a bandwidth of about 350 GHz) loses many of its high frequency components and broadens to several picoseconds after propagation of about 700 μm on a 50 μm diameter STM tip.

Second, a metal wire can act as an antenna and pick up radiation from the sample which does not necessarily tunnel through the junction. Antenna effects might be expected to persist unless the tip is much shorter than the typical wavelength of the radiation. The shortest wavelength in a one picosecond pulse propagating at the speed of light is on the order of 800 μm, so again we find that a good tip should be 100 μm or shorter.

Third, a long metal wire may produce reflections which interfere with the main signal. If we wish to have reflections separated by substantially less than the one picosecond resolution goal, we need a tip shorter than 100 μm. Finally, if the distance from the tip to the photoconductive switch is much shorter than the shortest wavelength in our signal, we expect to be able to model the tip assembly as a lumped circuit rather than as a nonlinear transmission line. This should substantially simplify the interpretation of our measurements.

There are a variety of lithographic techniques which generate short sharp tips on Si structures. A very small subset of these techniques generates metal tips. There is only one technique for forming metal tips compatible with SOS technology. This is the so-called "Spindt" process[4], in which metal is evaporated through an undercut hole in a layer on the sample surface. During evaporation, metal continually accumulates on the periphery of the hole, reducing the aperture size, so that the metal evaporated through the hole forms a cone. The transition metals titanium and niobium, each of which produces a surface oxide in air, produce the sharpest cones.

The technique has been used to make AFM tips[5], in which case a surface oxide does not affect the operation of the instrument, and to make contact probes[6], in which case the oxide is broken during contact. Our attempts to use this technique with sputtered aluminum were not successful and produced cylinders instead of cones. The most

likely cause is that the source was not highly directional due to the size of the sputtered target and its proximity to the sample.

In STM, the choice of tip material is limited, and it is not clear that appropriate tips can be developed for use in air with such a Spindt evaporation. For example, platinum-rhodium alloy, the material of choice for STM tips, must be sputtered to preserve its composition. Ultimately the addition of a reliable Spindt evaporation step to our process should give us a reproducible source of mass produced tip assemblies. In the absence of this capability, we have used less aesthetic but highly effective techniques to attach tips.

Initially, we tried to make short tips using a wire bonder. After completing a bond, the bonder advances its wire by a very small amount, approximately 100 μm. When the next bond is made, this extra wire ends up sticking out of the back end of the bond at a 10 to 30 degree angle from the vertical. The parameters of the bonder can be set so that the bonding tip pulls up too quickly to leave a real bond but pulls slowly enough that it leaves the extra wire. The length of this "wire-bond" tip can be increased reproducibly by using the thread function of the bonder.

There are several disadvantages to "wire-bond" tips. For one, the choice of metal is limited to aluminum or gold. Additionally, we used a wedge bonder, so the wire is flattened at the point of contact with the surface. In practice, this flattened portion acts as a soft spring and makes stable tunneling difficult. Typically, the wire is sucked into the surface by Van der Waals forces and springs away only after the piezo has retracted by many tens of angstroms. Hence the tip can never stably hover in tunneling range several angstroms from the surface.

It is probably possible to strengthen "wire-bond" tips with a bit of epoxy at the flattened portion. It may also be possible to modify the bonder to enable the use of Pt-Rh wire. However, we decided to adopt an even simpler method for attaching tips.

First, the sapphire substrate is diced into individual chips, each containing a photoconductive switch as close to the edge of the chip as possible. Then we cut a thin wire to a length of 30 μm - 500 μm with a sharp blade under an optical microscope. Before it is cut, the wire optionally can be electrochemically etched to produce a very sharp and well-defined radius of curvature at the of the tip. The wire is glued with Epotek H60E conductive silver epoxy as close to the switch as possible. The method does not work if the contacts to the switch are aluminum, because the epoxy does not penetrate the oxide layer. The first tip we made with this technique consisted of a 700 μm long 50 μm diameter brass wire attached to the end of a 1 x 10 mm rectangular beam. In all successive iterations, we used 25 μm diameter 90/10 Pt-Rh wire.

The procedure is straightforward and simple when the wire is 200 μm long or longer. In this case, the wire can be moved and placed with a pair of fine tweezers. The wire is glued to and is supported by the edge of the chip, and a trace of conductive epoxy is drawn from the wire to the metal contact on the switch. In order to bring the wire as close as possible to the switch, the end of the chip is mechanically polished with a diamond file until the metal contact extends to the edge.

The wire is often not long enough to be braced against the edge of the chip. Instead, a small drop of epoxy is placed directly on the metal contact with a fine wire attached to an X-Y-Z stage with micron resolution. The tip is picked up by static electricity with a fine piece of sable fur or by suction at the end of a pulled capillary tube. The tip is brought to the epoxy ball with a similar X-Y-Z stage and placed on the epoxy. Once the tip breaks through the surface of the epoxy, surface tension pulls the tip away from the sable fur or capillary tube. The entire process is performed under a stereo optical microscope. Figure 2.1 illustrates the mounting geometries for short and long tips. Figure 2.2a shows a photograph looking down on a 35 μm tip assembled in this manner. Figure 2.2b shows an SEM image of a 20 μm tip and the photoconductive switch at the end of the chip.

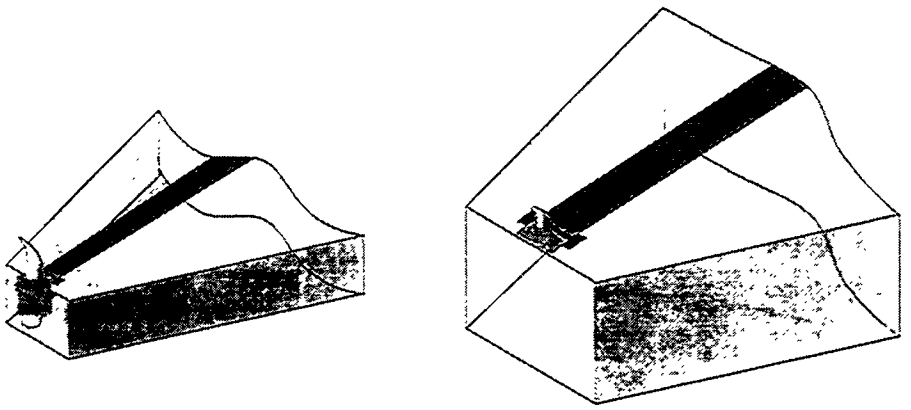


Figure 2.1. Mounting geometries for cut wire tips. **a)** A long tip (>200 μm) must be is braced on the edge of the chip for support. An epoxy trace makes electrical contact to the switch. **b)** A short tip (< 100 μm) must be mounted directly on the metal contact in a small pool of epoxy.

The shortest tip that can be cut and attached in a finite amount of time with the above technique is approximately 25 μm. When the tip is shorter than this length, it is difficult to distinguish from dust. In addition, the epoxy on the surface often contaminates the butter side of the tip. Therefore, to make shorter tips we borrow a technique developed in Germany[7] for creating tips from silver epoxy.

The method is to apply a small drop of silver epoxy, to begin to cure it, and to insert and remove an extremely sharp object like a pulled optical fiber when the epoxy has hardened to an appropriate consistency. The epoxy sticks to the fiber and forms a sharp cone as the fiber is withdrawn. We found that our epoxy does not stick well to optical fibers. Instead of using a fiber, we place a small ball of epoxy at the end of a fine wire. We touch the epoxy on the wire to the epoxy on the surface and gradually separate them. A narrow neck of epoxy results, and further separation breaks the neck into two sharp opposing points. We view epoxy on the surface from the side to verify that it is sharp and then heat-cure it. The technique works best with old epoxy which has already partially hardened.

A major advantage of this method is that the tip can be placed within 10 μm of the switch. The length of the tip also can be varied (not reproducibly) from about 5 μm to 100 μm. The disadvantage of this method is that the tip is not extremely sharp; its radius of curvature is usually on the order of 1 μm. Also, even after curing, the epoxy is much softer than a Pt-Rh tip, and tends to deform when "crashed" into the sample.

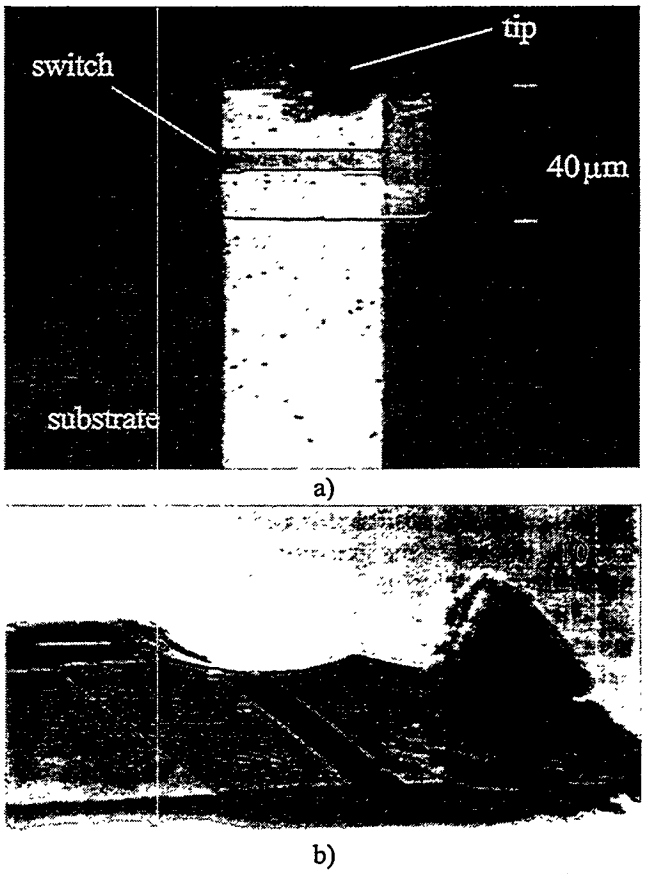


Figure 2.2. Short Pt-Rh tips. a) An optical microscope image of a 35 μm Pt-Rh tip attached with conductive epoxy. The total distance from the point of the tip to the switch is 55 microns. The corner of the substrate was polished to bring the switch closer to the end of the chip. b) An SEM image of a 20 μm Pt-Rh tip assembled by Jennifer Glass.

When the tip is very short, 30 μm or less, it is possible for the chip to contact or scratch the sample surface. To prevent this, the tip assembly normally is held at a slight angle to ensure that most of the chip's surface is far from the sample surface. However, if the tip is not exactly at the edge of the chip, then the edge can extend below the level of the tip itself. A more severe problem is accidental sideways tilt in the tip or the sample. On a 1 mm wide beam, six degrees of tilt will lower the corner by 50 μm.

To address these concerns, the tip is placed as close as possible to the end of the beam. With the dicing saw, the closest the edge can be cut to the switch is about 30 μm. Therefore the end of the chip is carefully polished with a diamond file to bring the switch within 10 μm of the edge.

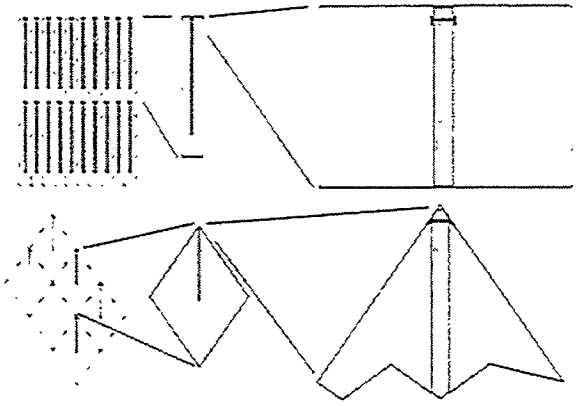


Figure 2.3. Two different ways of generating an array of switches and dicing them into chips. The diamond shaped chips are more suitable for use with very short tips because it is more difficult to accidentally contact the sample with the substrate.

Additionally, the end of the chip must be cut to be as narrow as possible. A simple way to do this is to create an array of tips on a diamond shaped rather than rectangular lattice. Figure 2.3 illustrates these two options. The polishing which brings the switch close to the end of the tip leaves the flat end region on which to mount long tips illustrated in Figure 2.1.

With the development of diamond-shaped chips, we were tempted to try one final technique to generate a short tip. We cut the chip with a dicing saw so that the corner of the diamond tip was covered by the metal contact to the switch. When the chip was held at an angle, the bottom corner, coated with metal, formed a sharp tip. This method generates tips with effective lengths between 10 μm and 50

μm with a yield of about 20%.

These tips make reasonable STM images, but the image quality is not comparable to what is obtained with our Pt-Rh wires. Also, when the tip is crashed several times, the metal at the end is removed and the tip can no longer tunnel or produce images. Even if the tip is not crashed, it gradually degrades with repeated imaging and the tunnel gap becomes unstable. We decided that these tips were unsuitable compared to Pt-Rh wire tips for imaging.

There are several additional difficulties with the diced tips. It is difficult to determine if the tip will tunnel without actually trying, because the shallow angle of incidence makes it difficult to judge the position of the lowest point with an accuracy better than ±10 μm. Often a crumb of sapphire protrudes further than any of the metal and may scratch and destroy part of the sample upon approach. Even if the tip does function properly, it still can be damaged during approach, since it is only on the order of 3000 Å thick.

2.3. STM Heads

The STM head must provide optical access to the tip and sample. It must be compact and rigid to resist noise from the laser cooling water, moving time delay stages, and other sources. Our initial experiments, in the Spring and Summer of 1993, were performed with a STM which we borrowed and modified for ultrafast work. Deficiencies in the design of this head inspired the second generation head. Each head is described below.

2.3.1. Original STM

The original STM, illustrated in Figure 2.4a, consists of a brass top plate clamped with springs to three 80 turns-per-inch height adjustment screws on a cylindrical main body. The screws have stainless steel ball bearings at their ends. Two screws contact the plate approximately one millimeter off a lateral line through its center. The bearing on one screw fits in a machined dimple on the bottom surface of the plate which acts as a fixed pivot point. The second screw sits in a V-shaped groove machined parallel to the line joining the two screws, and it is free to move along the groove as the relative height of the screws changes. These screws provide a coarse vertical adjustment. The third screw rests on the flat surface of the plate about 40 millimeters off axis and provides fine vertical adjustment at the center of the plate by using the two off axis screws as a fulcrum.

The top plate has a one inch diameter hole through its center. The chip containing the tip is glued to a 16 pin chip holder which has been cut in half width-wise. This tip assembly plugs into a 16 pin socket mounted 15 degrees from horizontal on a brass holder. The end of the tip assembly extends about two thirds of the way through the hole in the plate. A bolt is screwed through a channel in the holder to the top plate. The holder can be moved laterally to permit coarse positioning of the tip relative to the sample. The geometry of the top plate, is shown in Figure 2.4b.

The approach is accomplished first with the near-axis screws and at the end with the off-axis screw. As the height changes, the tip moves laterally with respect to the sample. In order to land on a particular spot, the tip holder must be unscrewed from the top plate and repositioned several times during the approach. Finer lateral positioning is accomplished by "walking" the various approach screws, adjusting one, and compensating for the resulting lateral motion with the others.

A piezo tube scanner[8] enables lateral motion and regulates the separation between tip and sample. A stainless steel base and a magnetic steel cap with a small samarium-cobalt magnet in it are soldered to the piezo. (The base and cap can also be glued with a rigid epoxy like Epotek H21) The steel base is bolted to the bottom of a cylindrical bore through the main body of the microscope so that the piezo tube comes up through the bore. The top of the piezo's cap is flush with the top surface of the main body. The space between the bore hole and the walls of the piezo tube is filled with a silicone rubber glue which protects the piezo from damage

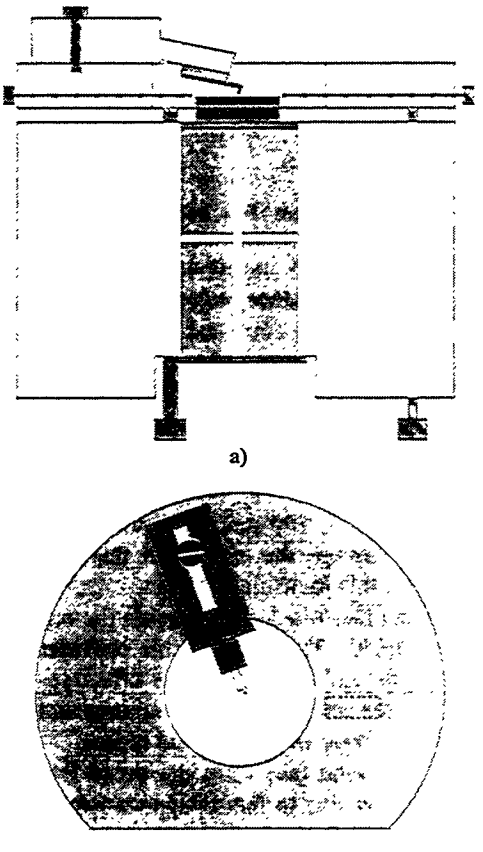


Figure 2.4. A schematic view of the original microscope used for USTM experiments. **a)** The base and top plate (the springs which clamp the top plate to the main base are not shown.). **b)** The top plate and tip assembly.

but does not restrict its motion.

The sample is mounted in a 16 pin chip holder and is plugged into a 16 pin socket glued to a magnetic steel disc. The steel disc is held magnetically by the samarium-cobalt magnet in the piezo cap. In this arrangement, the top of the sample is approximately 0.3 inches above the surface of the main body.

The piezo tube is 1.414 inches long, 0.5 inches in diameter, and has 0.020 inch thick walls. It is formed from PZT-5 ceramic poled in the outward radial direction and is plated with nickel inside and out. The outer nickel layer is separated into four electrically isolated quadrants by removing thin vertical strips of the nickel coating and is separated into upper and lower sections by removing a radial strip of the coating midway up the tube. We mentally label the sections ±X Offset, ±X Scan, ±Y Offset, ±Y Scan, and Z. The Z-electrode is the inside nickel coating. The scan electrodes are used to generate the x-y raster scan during imaging. The offset electrodes are used for fine lateral positioning. An illustration of the sectioning and the piezo tube appears in Figure 2.5.

One thin Teflon coated wire is soldered to each of the eight electrodes and to the inside coating of the tube. The wires are connected to separate high voltage supplies in the STM electronics. In order to produce lateral motion, a positive voltage is applied to one electrode of a diametrically opposed pair while an equal amplitude negative voltage is applied to the other (the voltages are referenced to the potential of the Z electrode). The tube bends opposite the direction of the voltage gradient as the thickness of the high voltage side increases (so that its length shrinks) and the thickness of the low voltage side decreases. When a positive voltage is applied uniformly to the outside electrodes relative to the Z electrode, the tube wall expands and the height of the tube shrinks. The amount of motion per volt, determined empirically, is 125 A/V in the X and Y directions and 60 A/V in the Z direction. The range of motion, defined by the ±130 Volt range of the high voltage outputs, is approximately 3 μm in the X and Y directions and 1.5 μm in the Z direction.

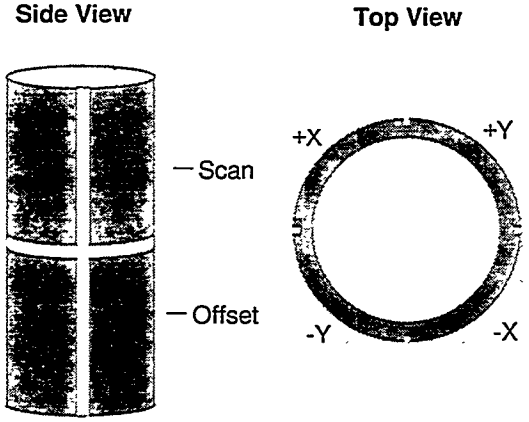


Figure 2.5. The pattern of electrodes on the piezo tube.

The microscope is clamped on its side in a heavy brass block. The magnetic holder and the springs which clamp the top plate to the main body are strong enough that the microscope can be operated stably in this orientation. The photographs in Figure 2.6 illustrate the geometry of the STM. A single long working distance microscope objective (Mitutoyo M Plan APO 10) focuses both laser beams along a horizontal path to the microscope and brings one beam to the tip and one to the sample. The same objective and an external lens provide an image of the top of the tip assembly and sample on a TV monitor connected to a CCD camera. The TV imaging system is used during the approach and allows an accuracy of a few μm in positioning the tip relative to the sample.

The use of a single objective for both beams restricts the lateral separation between the beams to be less than about 500 μm. It also makes it difficult to optimally focus both beams, since the switch on the tip assembly can be on the order of 100 μm above the surface of the sample. The fact that the pump laser beam on the sample travels entirely through the air, while the probe laser beam on the tip focuses through the sapphire substrate slightly alleviates this focusing problem.

The brass block which holds the microscope sits on a hard foam pad to isolate the STM from vibrations on the optical table. The optical table provides additional vibration isolation from the building. The hard foam pad was chosen instead of other more effective isolation techniques (such as metal plates stacked on Viton dampers or hanging the microscope from springs and using eddy current damping) in order to minimize relative motion between the microscope and the optical beams.

The connection from the tip to the pre-amplifier input is kept as short as possible to minimize noise in the preamplifier due to input capacitance; its length is typically on the order of 10 cm. A box wrapped with grounded aluminum foil surrounds the microscope and provides electrical and acoustical shielding.

The STM achieves atomic resolution on highly oriented pyrolytic graphite (HOPG) and on TaS_2 (a layered material exhibiting charge density waves) in air. It easily attains several nanometer lateral resolution on gold. Conventional STM images of these surfaces appear in Figure 2.7.

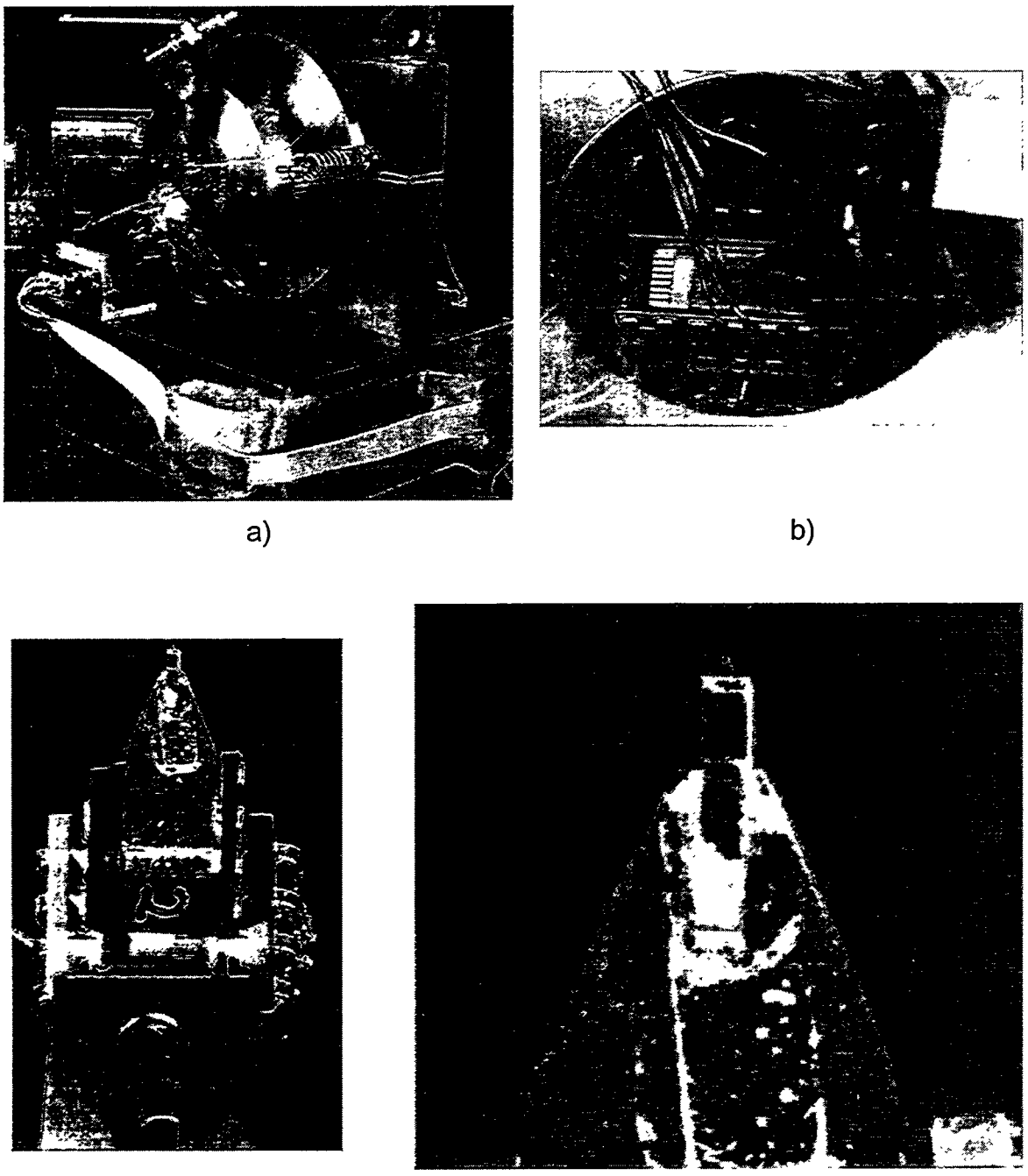


Figure 2.6. Photographs of the original USTM head. a) The head clamped into its brass holder. The focusing objective is visible at the top left of the image. b) A closer view of the tip and a sample. c) The tip assembly and tip holder. d) A closer view of the first tip we used. The wire at the top is 700 microns long, 2 mil diameter brass.

2.3.2. Deficiencies of the original STM

There were several inconvenient features in our original STM which we have attempted to eliminate with an improved design. In the original STM, the tip assembly was held at an angle of 15 degrees from horizontal. This made it extremely difficult to view the tip from above the sapphire substrate unless it was bent to project forward from the chip edge. The requirement of a bent or projected tip limited us to tips of length 200 μm and longer.

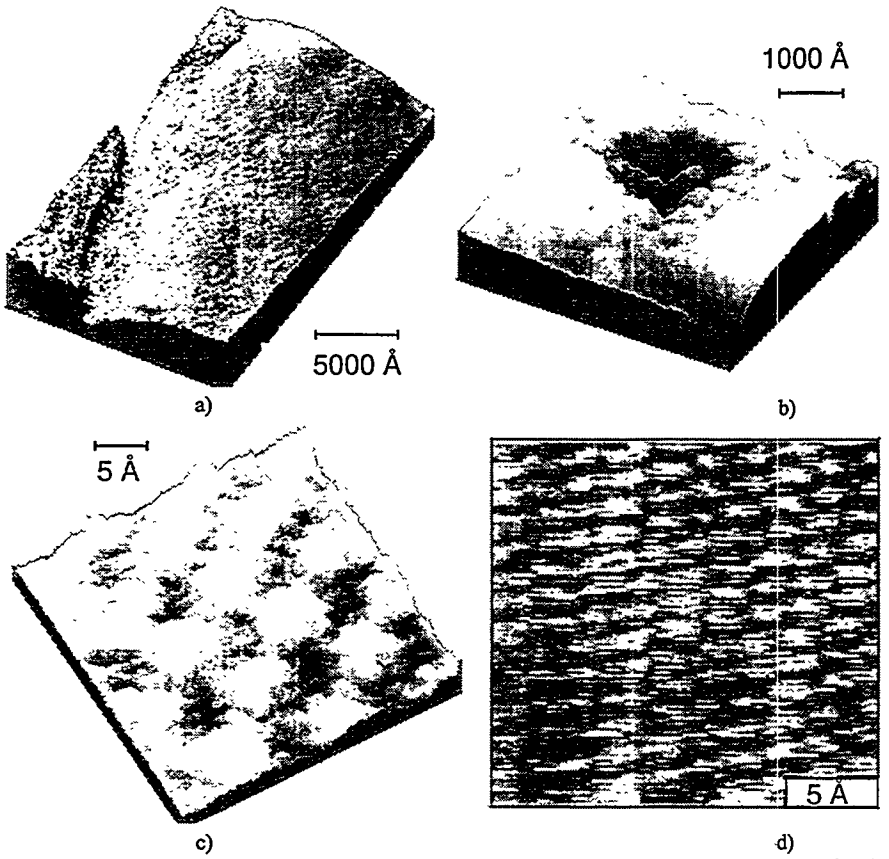


Figure 2.7. Some representative images measured with the original head. All except for the second are raw data. a) A large area scan of an aluminum transmission line deposited on sapphire. b) A 0.7 x 0.7 micron scan of a gold transmission line on sapphire. The dimple was caused by crashing the tip (accidentally) into the sample. c) A charge density wave (CDW) on 1T - TaS_2. The period of the CDW is 12.4 Angstroms. The atomic lattice is visible as a modulation of the CDW. d) Atomic resolution on HOPG. The period is 2.5 Å.

The lateral position of the tip assembly was very difficult to adjust on a micron scale. In the experiments discussed in Chapters 3 and 4, it was necessary to land on a 10 μm wide transmission line on the sample. In order to accomplish this, we needed to walk the height adjustment screws. The motion of the tip in the z-direction was then coupled to its transverse motion. Also, the smallest increment in height we could achieve during the approach was dictated by the resolution of the coarse adjustment screws, and it was difficult to avoid hard crashes to the surface.

The objective which coupled the laser beams to the STM was not rigidly attached to the body of the STM. This permitted relative motion between tip, sample, and beams and therefore introduced noise to the measurement from fluctuations in the conductance of the switch on the tip.

The tip-sample area was not optically accessible over a large solid angle. The lack of a side-view imaging system made it impossible to align to and approach the sample with very short tips. The versatility of the instrument also was limited by the lack of optical access. Beams could not be brought to the side of a sample in order to bring laser light directly beneath the tip, and beams could not be brought in from beneath the sample to excite thin films.

2.3.3. The Second Generation STM

The second generation STM, schematically illustrated in Figure 2.8, was designed to address the limitations discussed above and to be a more stable and versatile system than the original microscope. It is based on a three piezo inertial walker design by Besocke[9].

The heart of the microscope is a stainless steel ring with three 120 degree helical regions on its bottom, illustrated in Figure 2.8 and upside down in Figure 2.9b. The ring rests on three sapphire ball bearings on top of three piezo tubes located at 120 degree intervals on a circle around the samples. The tip assembly is mounted to the ring.

The piezos are prepared in almost the same manner as in the original STM[10], and are soldered to an aluminum ring which is bolted to the microscope

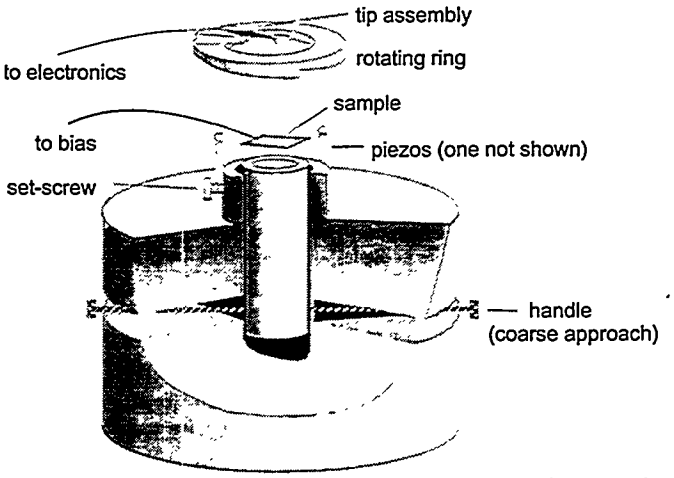


Figure 2.8. The second generation STM, based on the three piezo Besocke design. There is optical access to the sample over a nearly 2π solid angle from the top and over a small area from directly below.

base. They are 0.5 inch long, 0.125 inch in diameter, and have 0.020 inch walls. Their displacement with applied voltage is 150 Å/V in the X and Y directions and 20 Å/V in the Z direction. The sapphire bearings are glued on with Epotek H21 epoxy.

During a conventional image scan, the piezos work in concert. Identical voltages are applied to the equivalent electrodes on each of the three piezos. If the scan is sufficiently slow, static friction between the ring and the ball bearings prevents the ring from sliding, and the microscope behaves as if it were a single tube scanner. If the scanning speed is too great, the force at the turning points becomes sufficient to break the force of static friction and allow the ring to slide.

Although a sliding ring is not amenable to STM imaging, this inertial motion can be harnessed and used to provide vertical and lateral translation. The ingredients in the process are illustrated in Figure 2.9. The voltage waveform shown in Figure 2.9a, consisting of a slow ramp up and a sudden drop to zero, causes the piezos to gradually move in one direction and then to snap back suddenly. When appropriate parameters for the rise and decay of a waveform are chosen, the ring follows the slow motion and then remains in place while the piezos return to their original positions.

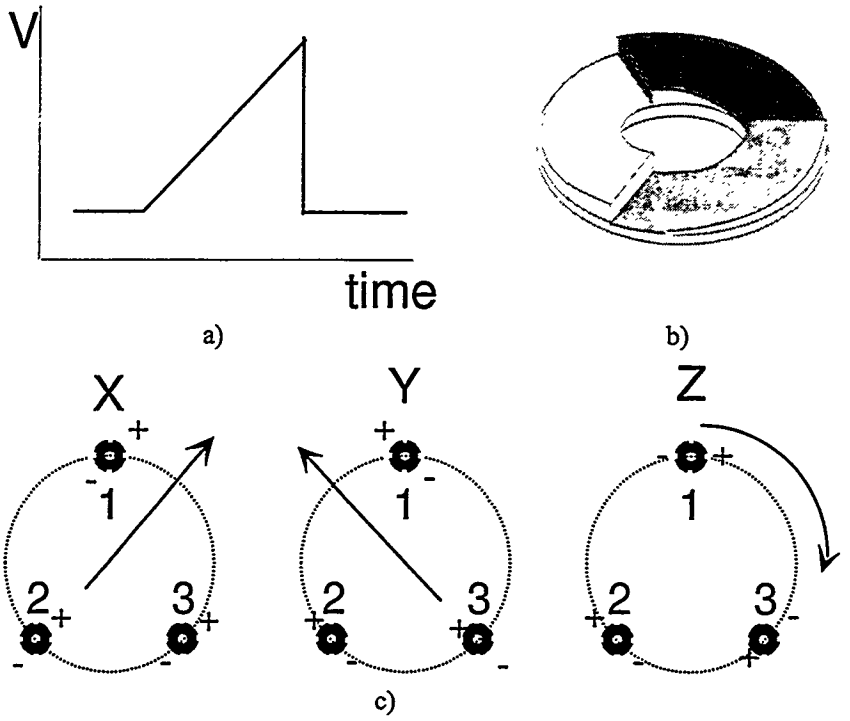


Figure 2.9. Elements of the inertial motion positioning system **a)** A voltage waveform used to produce inertial motion. A typical peak amplitude is 20 volts. The rise time is on the order of 1-2 ms, while the fall time is on the order of 20 μs. The spacing between pulses in a train is immaterial as long as the ring has a few ms to settle. **b)** A ring with three helical ramps permits motion in the vertical direction. **c)** The direction voltage should be applied to the piezos to produce the indicated motion of the ring.

A train of voltage waveforms causes the ring to move in steps in the desired direction. The voltage is applied identically either to the X electrodes or the Y electrodes of each piezo to produce X or Y motion. To produce vertical motion, the voltage is applied vectorially to the X and Y electrodes of each piezo so that each piezo moves in a direction tangent to the circle which contains their centers. This causes the ring to rotate along its ramps.

The inertial motion can produce steps ranging from several hundred angstroms out to one micron. The exact direction of motion depends on the presence of contaminants on the ring and piezo and on how perfectly horizontally the ring sits. Hence X and Y motion is not perfectly orthogonal, and rotational Z motion is usually coupled to some linear translation. Occasionally, the ring gets stuck and must be jostled back and forth by flipping the polarity of the driving waveform to break it free.

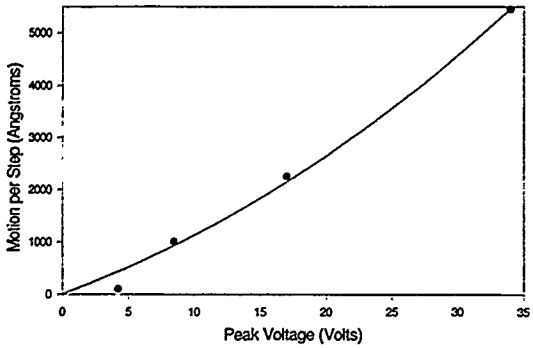


Figure 2.10. The variation of average motion per inertial step with peak voltage of the driving waveform. The waveform risetime is 1.8 ms, and the fall time is 20 us. The variation in the average motion per step in repeated trials of 100 steps is approximately 5%. The solid line is a second order regression forced to pass through (0,0).

The peak and width of the voltage waveform which produces reproducible steps in a given direction must be determined empirically. In our STM, a peak of 34 V, a rise time of 1.8 ms, and a fall time of 20 us[11] yields reproducible 5500 Å steps in the X or Y direction with about 5% variability from step to step. The spacing between pulses in a train is immaterial as long as the ring has a few ms to settle.

The step length scales superlinearly with peak voltage. This is graphically illustrated in Figure 2.10, which shows the variation in average step length with peak voltage. Below 1000 Å per step, the average motion per step appears to drop precipitously. This drop most likely results from missed steps

which occur because the motion of the piezos is no longer jarring enough to consistently dislodge the ring.

Circumferential motion which rotates the ring downwards is nearly as reproducible as lateral motion. Although the minimum reproducible lateral step we can make is slightly less than 1000 Å, the minimum reproducible step in height, which depends on the slope of the helices, is substantially smaller. Our ring has 3 degree sloped helices, so the minimum vertical step is on the order of 100 Å. This small step size enables the tip to smoothly approach the sample with little danger of crashing.

Circumferential motion which rotates the ring upwards is extremely irreproducible. Unfortunately, there is a very small set of parameters for the voltage waveform which will produce this motion, and the length of the resulting steps is usually very short. The slightest bit of dirt usually causes the ring to translate instead of rotating. These deficiencies are not extremely important, because the design of the head allows us to push the sample out of range manually.

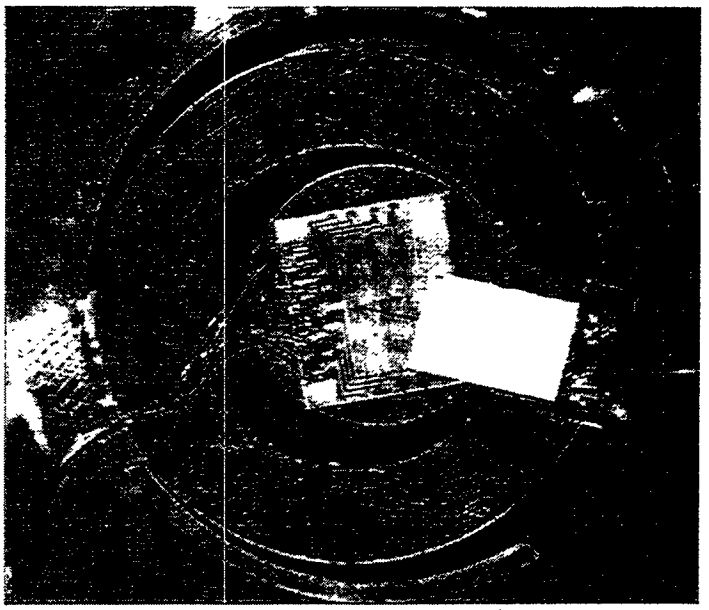


Figure 2.11. A top view photograph of the tip and sample

The sample is mounted magnetically to the top of a hollow cylinder which can slide or rotate inside a cylindrical hole in the main body of the STM to provide coarse motion of the sample relative to the tip. Two handles which attach to the cylinder through opposing 30 degree 0.5 inch high arcs cut out of the main body facilitate adjustment of the cylinder. A spring-loaded screw holds the cylinder firmly in place but permits alignment. The cylindrical hole is double-bored opposite the screw to provide two lines which give the cylinder additional stability when the screw presses against it.

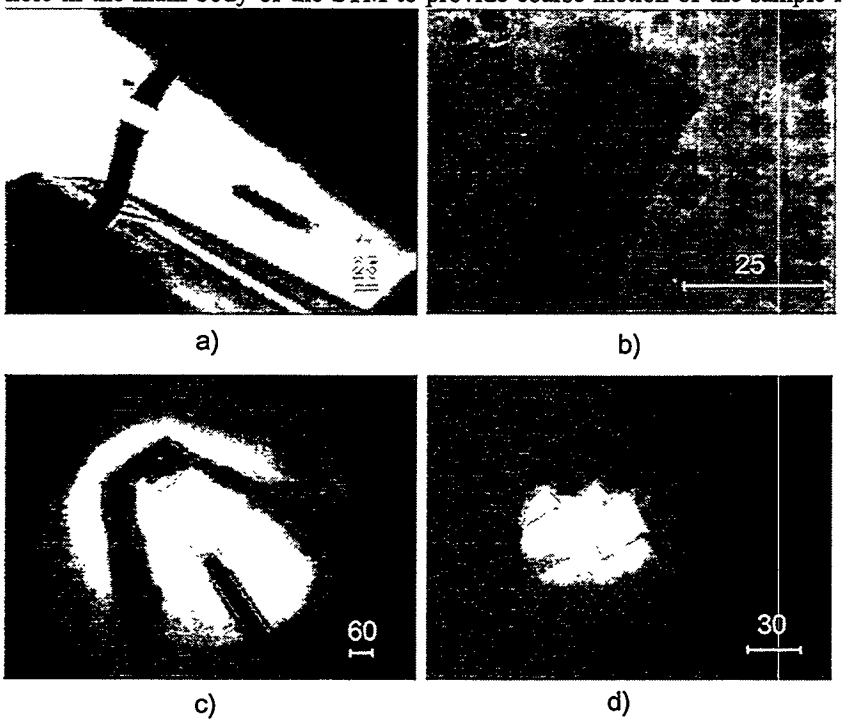


Figure 2.12. Photographs of typical views with the side and top imaging objectives. Scales are in microns **a)** A low magnification side view of a tip hovering over a gold surface. The actual surface is halfway between the real tip (above) and its reflection (below). The plane of the surface in the image tilts up from right to left. **b)** A magnified view of the same region. The magnification is accomplished by changing from a 10x to a 50x objective. **c)** Top view of the tip over the same gold surface, which is out of focus. **d)** A magnified view from the top. The magnification is accomplished by removing the second lens in the system and using the objective as a single lens magnifier. The double image, which also exists in "c)", arises from a second reflection on the microscope slide used to couple the image to the CCD.

A diamond shaped chip containing the tip is glued to a rectangular ceramic mount which is clamped to a raised step on the rotating ring. The step is tilted towards the center of the ring at an angle of 3 degrees. The length of the chip and ceramic assembly is such that the tip is within 1 mm of the center axis of the ring. The closer the tip is to the center of the ring, the less it moves laterally when the ring rotates. The ring has an outer diameter of 1.5 inches, an inner diameter of 0.95 inches, and a helical ramp height of 1.5 mm. It therefore has a lateral range of ±0.125 inches and a vertical range of 1.5 mm. A cylindrical aluminum shield bolted to

the microscope body prevents the ring from traveling far enough to fall off the piezos and prevents the user from breaking the piezos accidentally.

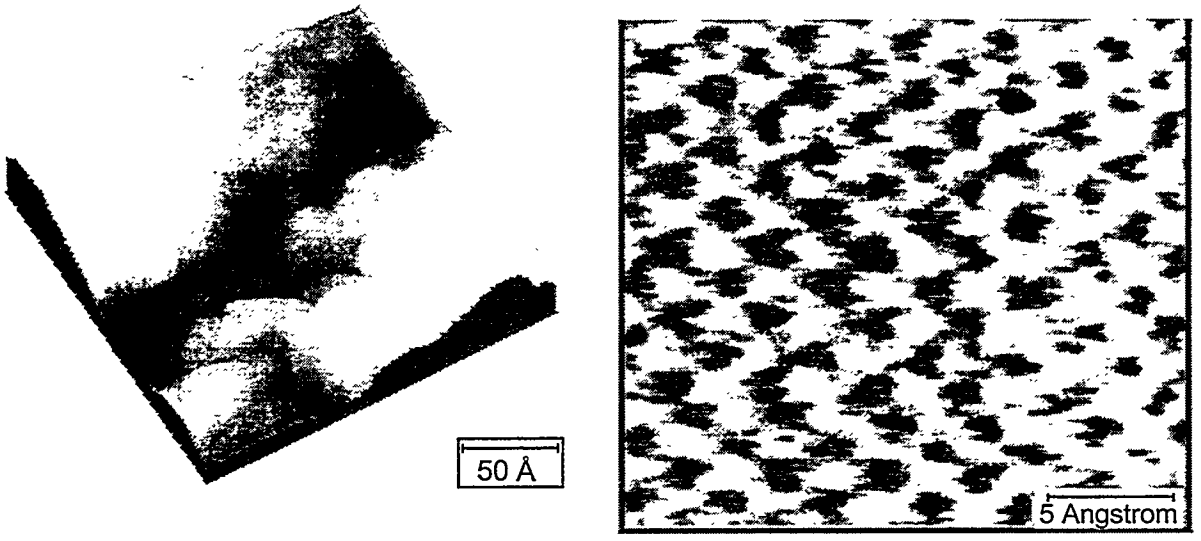
The tip meets the sample approximately 0.15 inches above the top of the ring, so there is optical access over nearly a 2π solid angle above the tip. The photograph of the tip and sample in Figure 2.11 illustrates this optical accessibility. A freely positionable objective inside the hollow sliding cylinder permits access from a smaller solid angle below the sample.

Two Mitutoyo long working distance objectives are available to bring light to the tip and sample from the top or from the side, and each objective also provides an image of the tip and sample. Typical images from the top and side are shown in Figure 2.12.

The side-view imaging system, with a maximum resolution of about 5 μm, permits us to carefully bring the sample close to the tip and permits the use of extremely short tips. The soft friction from the spring loaded screw enables micron scale height adjustment by hand. With practice, it is possible to reliably bring the tip within 3 μm of the sample. At this range, the approach via inertial motion is fairly trivial.

The probe laser beam is always focused to the tip through the top-view objective. In order to minimize relative motion between the laser beam and the STM tip, the top-view objective is attached to a sturdy X-Y-Z stage rigidly bolted to a plate which rests kinematically on ball bearings on top of three posts screwed into the microscope body. Relative motion between the laser beams and the objective is demagnified by approximately the ratio of the diameter of the focused beam to the diameter of the beam when it hits the objective. This relative motion can arise from beam wobble as the time delay is scanned, movement of the microscope on its vibration isolation platform, or imperfect laser pointing stability, which also adds angular fluctuations which are not demagnified. The microscope sits on a stiff foam pad which provides some vibration isolation without much allowing too much motion of the STM relative to the table.

The wire from the tip to the preamplifier is approximately 3 cm long, much shorter than in the original microscope. A grounded electrical shield around the microscope seems to have little effect on the noise performance. Figure 2.13 shows typical images made with the new microscope. The performance and stability of the new microscope is better than the original.



a) b)

Figure 2.13. Typical images collected with the new microscope. a) A 300 x 300 Å region on gold evaporated on mica. b) Atomic resolution on HOPG.

2.4. STM Electronics

The STM electronics consists of: a current preamplifier to convert nanoampere scale currents to measurable voltage levels; a feedback system which attempts to fix the tunnel current by varying the separation between tip and sample; voltage ramp and step generators to produce raster scans and a fast A/D converter to acquire image data; and image collection and analysis software. We use commercial products, the RHK Technology model STM 100 head electronics and the RHK Technology STIMAGE386 acquisition software, in our system. The feedback loop consists of proportional and integrating stages. The software runs on an IBM compatible PC based on the Intel 486 CPU and

controls a DT2821 250 kS/s A/D card from Data Translation.

2.5. Optics

2.5.1. Laser Source

In order to obtain good signal to noise levels and high temporal resolution in photoconductive sampling measurements, it is important to have a quiet, reliable laser source which produces optical pulses much shorter than the photoconductive switching time. We use a mode-locked Coherent Mira 900 Ti:sapphire laser pumped by a Coherent Innova 310 Argon laser. The Mira, pumped at 8 Watts, produces 700-800 mW of 120 fs pulses at a 76 Mhz repetition rate. We usually tune the laser to the peak of its gain, approximately 780 nm, but the performance of the switches does not seem to be affected much by laser wavelength between 700 and 800 nm. The beams have average powers of up to 300 mW each at the tip and sample.

2.5.2. Beam path[12]

Figure 2.14 illustrates of the circuitous routes of the laser beams to the STM. Pulses from the laser are split into a pump and a probe beam which travel separate paths of nearly equal lengths. Each path contains a variable time-delay. On one path, a computer controlled stepper motor permits a change in path length of up to 20 mm, corresponding to 130 ps of time delay, in 0.1 μm steps. On the other path, a fast scanner consisting of a galvanometer coupled to a low mass linear stage permits path length changes on the order of 1 cm at frequencies from 100 mHz to more than 50 Hz. Usually, only one translator or the other is used during a given experiment.

A small portion of each beam is diverted to a doubling crystal and a photomultiplier tube to provide a real-time monitoring system which operates when data is taken with the fast scanner. The beams which leave the autocorrelator have traveled equal path lengths and are therefore at zero relative time delay. A compensating translation stage allows us to adjust the time delay downstream without disturbing the autocorrelator. The retroreflector on the compensation stage is attached to a horizontal shaker which can move on the order of 5 μm at frequencies greater than 1 kHz.

An 8 inch clear path after the compensation stage permits us to introduce optical components into the beams as required. We always have an adjustable attenuator on each beam. When we perform low frequency time-delay scans, a chopper modulates one or both of the beams. Sometimes it is helpful to have a telescope on one or both beams to control the spot size and the focal length at the STM. A retroreflector on a linear stage can be inserted kinematically to divert the pump beam to the side or bottom objective.

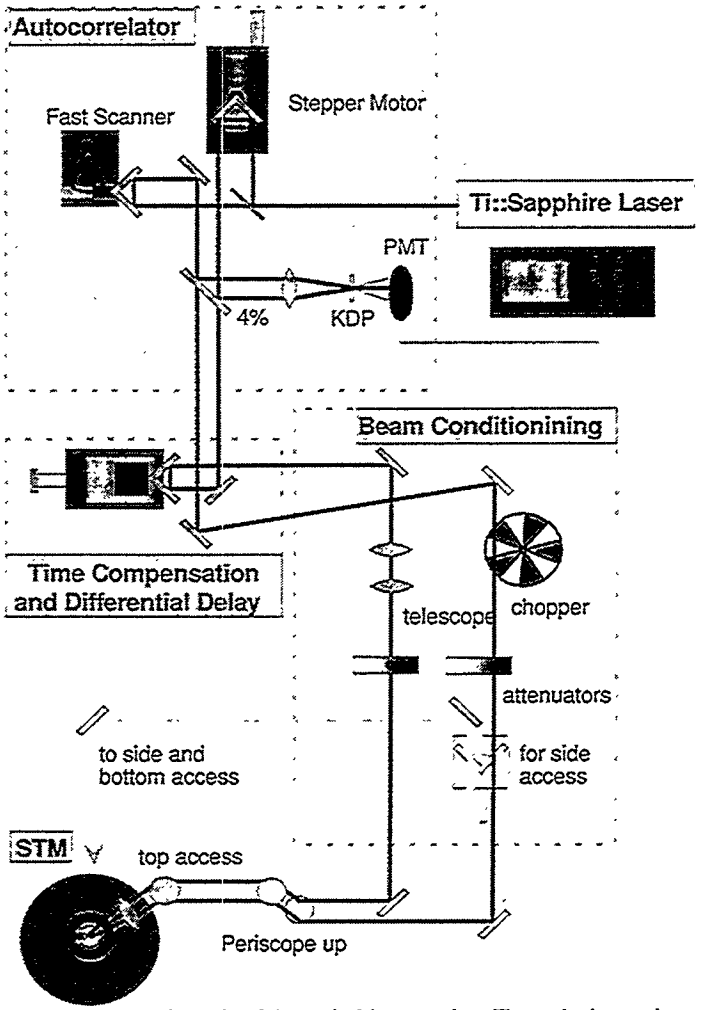


Figure 2.14. A schematic of the optical beam paths. The probe beam always hits the tip from above. The pump beam can be directed to the STM sample from the side or from below.

2.5.3. Noise Considerations and Scanning Techniques

In ultrafast spectroscopic measurements, the time delay is scanned, and the signal is measured and recorded as a function of time delay. In what I call a slow scan, the time delay is scanned at frequencies between 100 μHz and 1 Hz, the laser beams are chopped at frequencies ranging from several hundred Hz to several MHz, and lock-in

detection at the chopping frequency reduces the noise bandwidth to enhance the signal to noise ratio (SNR). The lock-in must pass enough bandwidth to collect the signal without distortion. For example, if a 100 second scan consists of 100 points, the lock-in must pass a bandwidth on the order of a few Hz for a signal with sharp features. Noise within this passed bandwidth can not be distinguished from the signal. Additionally, low frequency laser noise cannot be eliminated by chopping the laser beam.

Modern mode-locked lasers such as Ti:sapphire are extremely stable, so low frequency laser noise is usually not an important concern, but slow drifts in optical components may affect the signal in a similar way. In very sensitive ultrafast optical experiments where low frequency laser noise dominates the signal, researchers often use a fast-scanning technique with signal averaging in order to enhance the SNR[13].

In the fast scan technique, time delay is scanned at between 10 Hz and 500 Hz, and the signal is measured over a large bandwidth. During the time it takes to perform a slow scan measurement, hundreds or thousands of traces may be averaged together with the fast scan method. In a slow scan, low frequency noise appears as point to point scatter or smooth deformations of the true signal shape, while high frequency noise, except near the chopper frequency, is filtered away. On the other hand, in a fast scan, low frequency noise manifests itself as a normally unimportant scan to scan scatter in the DC baseline of the time-resolved signal. High frequency noise manifests itself as point to point scatter which averages to zero after many scans. Appendix 2 discusses the noise levels of the two scanning techniques in more detail.

Whether a fast scan will have a better SNR than a slow scan, given equal integration times, depends on the power spectrum of the noise in the experiment. Fast scanning is usually better than kHz mechanical chopping when the noise spectrum is dominated by 1/f and other low frequency noise[14]. In STM, low frequency mechanical noise and 1/f electrical noise are much more severe than in conventional ultrafast spectroscopy and constitute the dominant source of noise in the system. The best way to choose between scanning techniques is to do so empirically. In our system, fast scanning, even at the relatively low frequency of 20 Hz, usually yields a better SNR for an equal integration time than lock-in averaging with a slow scan.

Another advantage of fast-scanning is that we see the signal appear in real time and can rapidly respond to problems in the system. With a slow scan , we must wait until a substantial portion of the scan is complete in to get a sense of the quality of the data. Also, it is extremely satisfying to watch the signal evolve in a fast scan. Small signals magically rise from the noise as successive scans are averaged together.

A major disadvantage of fast scanning is that the scanner generates mechanical vibrations on the optical table. The signal which results from the coupling of these vibrations to the STM head is synchronized to the scanning frequency and therefore is not reduced in the signal averaging process. Our present sensitivity is limited by scanner vibrations. There are several techniques which provide rapid, vibration-free scanning.[15,16] A future improvement to the system will be the addition of a Heritage scanner[16], which uses very small angular motion of a mirror to produce up to 100 ps time delays.

2.5.4. Differential measurements

Another technique we can use to reduce noise is a differential time measurement, in which the time delay has an additional small modulation at a frequency on the order of 1 kHz. The shaker on the compensation stage in Figure 14 effects this modulation. The signal is sent to a lock-in with the shaker frequency as a reference, and the resulting signal is the derivative of the lock-in input signal. The input signal can be recovered by integration. With the proper choice of lock-in bandwidth, modulation frequency, and scanner frequency, this technique can improve the SNR by a small factor. With the LabCentral software discussed below, we can simultaneously measure the signal of interest and its derivative with this differential time modulation technique.

Since the resonant frequencies of the STM piezo tubes lie in the kHz range, the modulation frequency must be chosen with some care. I found that the shaker could produce a throw of 5 μm at 1.3 kHz without coupling noticeable vibrations to the STM. The improvement in the SNR depends on the bandwidth of the lock-in, which in turn fixes the maximum scanning frequency which permits an undistorted measurement. For the experiment described in the following chapter, the maximum scanner frequency for differential measurements was 9 Hz.

2.5.5. Scanners

We have used two different commercial fast scanners, each based on the same design[15]. A cam connected to a galvanometer shaft drives a linear slide with a thin stainless steel band. A hollow corner cube retroreflector is attached to the moving part of the slide. The first scanner, a Clark Instruments ODL-150, had a travel of almost an inch (150 ps) at up to 30 Hz. The second scanner, General Scanning's LT1320A, has a travel of only 1 cm, but it can scan at higher speeds while coupling fewer vibrations to the table than the Clark scanner. The LT1320A also

seems to degrade less with time than the Clark scanner. This degradation appears as increased vibration and wobble of the slider during a scan.

Each scanner is driven by an analog voltage waveform and provides a signal proportional to its instantaneous position. The motion is accurate to much better than 1% and reproducible to much better than the accuracy. The fast scanners can also be operated at frequencies between 0.1 Hz and 1 Hz, so they can be used in the slow scan technique. At frequencies below 0.75 Hz, the linear slide tends to move in a stick-slip fashion, so the motion is much less smooth and much less reproducible than motion at higher frequencies.

2.5.6. Software

The STIMAGE386 software is not designed to repetitively scan and average signals, so we developed our own software package, LabCentral, for this purpose. LabCentral is a Microsoft Windows® based program which controls a National Instruments AT-MIO16 A/D card and an AT-GPIB card. The program can generate a variety of waveforms to drive a scanner or can be triggered by an external oscillator which drives a scanner. It can simultaneously collect and average up to 8 traces, stored as arrays of points. During data acquisition, the user can display any number of graphs, each one containing any number of separate plots of one trace versus another or versus point number. The graphs can be updated separately in real time after the acquisition of each set of a specified number of scans.

The signals are averaged by adding the results of successive traces point by point. This averaging method is sufficient to produce 50 fs time resolution at scanning frequencies larger than 1 Hz, where the scanner motion is extremely reproducible. In principle, a binning scheme should be used for the averaging of scans at frequencies much below 1 Hz.

The program has sophisticated data analysis capabilities and a simple interface for the analysis - a calculator which manipulates arrays of data in the same way as it manipulates ordinary numbers. Analyses of stored data can be performed while new data is being collected, and the user can use other Windows programs while data is being collected. The program also has rudimentary printing capabilities. The interface appears in Appendix 3.

Chapter 2 References:

1. S. Gupta, M.Y. Frankel, J.A. Valdmanis, J.F. Whitaker, G. A. Mourou, F.W. Smith and A.R. Calawa, "Subpicosecond carrier lifetime in GaAs grown by molecular beam epitaxy at low temperatures", *Appl. Phys. Lett.*, **59**, p. 3276 (1991).
2. M.B. Ketchen, D. Grischkowsky, T.C. Chen, C-C. Chi, I.N. Duling, III, N.J. Halas, J-M. Halbout, J.A. Kash, and G.P. Li, "Generation of subpicosecond electrical pulses on coplanar transmission lines", *Appl. Phys. Lett.*, **48**, p. 752 (1986).
3. F.E. Doany, D. Grischkowsky, and C.-C. Chi, "Carrier lifetime versus ion-implantation dose in silicon on sapphire", *Appl. Phys. Lett.*, **50**, p. 460 (1987).
Our dose should produce a carrier lifetime of 600 ps.
4. C.A Spindt, I. Brodie, L. Humphrey, and E.R. Westerberg, "Physical properties of thin-film field emission cathodes with molybdenum cones", *J. Appl. Phys.*, 47, p.5248 (1976).
5. Thomas Robert Albrecht, "Advances in Atomic Force Microscopy and Scanning Tunneling Microscopy", Doctoral Thesis, Stanford University (1989).
F. Ho, A.S. Hou, D.M. Bloom, "High-speed integrated circuit probing using a scanning force microscope sampler", *Electron. Lett.*, **30**, p. 560 (1994) and "Picosecond electrical sampling using a scanning force microscope", *Electron. Lett.*, **28**, p. 2302 (1992).
6. Joungho Kim, Steven Williamson, John Nees, Shin-ichi Wakana, and John Whitaker, "Photoconductive sampling probe with 2.3-ps temporal resolution and 4-mV sensitivity", *Appl. Phys. Lett.*, **62**, p. 2268 (1993).
7. T. Pfeifer, H.- M. Heiliger, E. Stein von Kamienski; H.G. Roskos, and H. Kurz, "Fabrication and characterization of freely positionable silicon-on-sapphire photoconductive probes", *J. Opt. Soc. Amer. B*, **11**, p. 2547 (1994).
8. G. Binnig and D.P.E. Smith, "Single-tube three-dimensional scanner for scanning tunneling microscopy", *Rev. Sci. Inst.*, **57** , p.1688 (1986).
9. J. Frohn, J.F. Wolf, K. Besocke and M. Teske, "Coarse tip distance adjustment and positioner for a scanning tunneling microscope", *Rev. Sci. Instrum.*, **60**, 1200 (1989).
10. Each piezo is sectioned into four lengthwise electrodes and one z-electrode, but the electrodes are not separated radially into scan and offset halves. Instead, the scan and offset voltages are electrically summed and applied to the electrodes. The lack of a radial separator is relatively inconsequential in the design of the instrument. Piezos without a separator have slightly greater travel and are several dollars less expensive than their separated brethren.
11. The actual "snapping" time of the piezos is limited by their resonance frequencies for lateral motion. These are on the order of 4 kHz.
12. The optical layout in the original USTM system does not differ from the layout used with the new microscope in its essential details. The main difference is that beams must come vertically to the top in the new system, while they traversed a horizontal path in the old system. The detailed descriptions of the optical geometry in this chapter refer to the current geometry, used with the new STM head.
13. G.C. Cho, W. Kütt and H. Kurz, "Subpicosecond time-resolved coherent-phonon oscillations in GaAs", *Phys. Rev. Lett.*, **65**, 764 (1990).
14. D.C. Edelstein, R.B. Romney, R.B. and M. Scheuermann, "Rapid programmable 300 ps optical delay scanner and signal-averaging system for ultrafast measurements", *Rev. Sci. Inst.*, **62**, p. 579 (1991).
15. A. Black, R.B. Apte, and D.M. Bloom, "High-speed signal averaging system for periodic signals", *Rev. Sci. Instrum.*, **63**, p 3191 (1992).
16. K.F. Kwong, D. Yanklevich, K.C. Chu, J.P. Heritage, and A. Dienes, "400-Hz mechanical scanning optical delay line", *Opt. Lett.*, **18**, 558 (1993).

Chapter 3
IMPULSE RESPONSE OF THE USTM

3.1. Introduction

This chapter presents experimental measurements of the high frequency response of the tunneling junction of a scanning tunneling microscope in air. The measurement is performed in the time domain and is illustrated conceptually in Figure 3.1. The tunnel junction is formed between the tip of the STM and a metallic surface, and the lateral position of the tip is fixed. With a pump pulse and a photoconductive switch, we generate a short voltage pulse which propagates on the metal surface and transiently biases the junction. We measure the resulting signal as a function of time delay between the pump pulse and the probe pulse on the tip.

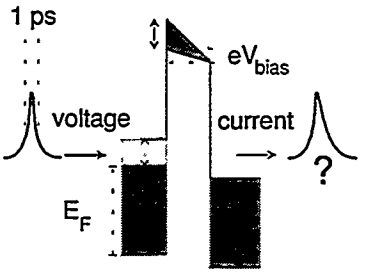


Figure 3.1. Principle of the experiment. A voltage pulse shakes the energy of the Fermi level on one side of the barrier on a picosecond time scale. The hatched regions show the effects of the pulse. The DC bias, V_{bias}, accounts for the offset between the Fermi levels on opposite sides of the barrier.

Without some knowledge of the shape of the driving voltage pulse, the measurement cannot provide very much information about the junction. Therefore, we use samples which permit us to characterize, independently from the STM, the voltage pulse which propagates on the transmission line. We also measure what I call the crashed signal, which results when the tip is in ohmic contact with the sample.

The experiment serves two purposes: it characterizes the response of the USTM, and it allows us to study the dynamic properties of the tunnel gap. A discussion of the latter requires some specific model for the USTM measurement and will be deferred until the next chapter. The primary focus of this chapter is the description of the mechanics of the experiment and the characterization of the USTM as an instrument. This characterization will give us a number of general empirical results which will allow us to test prospective models for the system. Our results also will permit us to determine the time resolution, sensitivity, and spatial resolution of the USTM based on general principles, without reference to any particular model.

3.2. Experimental setup

3.2.1. Samples

To avoid coherent artifacts and laser heating effects at the tip, it is wise to separate spatially the pump and probe beams. For this reason, we generate the voltage pulse on a high frequency coplanar transmission line and measure the signal after it has propagated several tens to hundreds of microns along the line. The samples are fabricated on SOS using the technique discussed in the previous chapter. Figure 3.2 illustrates a typical sample geometry and the experimental arrangement.

A typical transmission line consists of two 5 to 10 μm wide, several millimeters long Cr/Au or Al strips separated by 10 or 20 μm gaps. A single 60 x 60 μm Si pad bridges the gap between the lines. The voltage pulse is generated at this pad. A second pad bridges the gap between one of the lines and a third electrode which branches off to the side of the main transmission line. This second pad permits photoconductive sampling of the waveform generated on the sample[1].

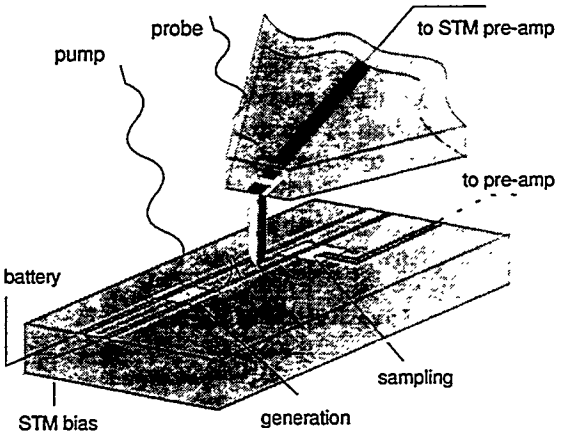


Figure 3.2. The experimental arrangement. A battery biases the generating Auston switch. A pump laser pulse strikes the switch to launch a voltage pulse which propagates along the transmission line to the STM tip. A probe laser pulse can be directed to the switch on the tip or to a sampling pad on the sample. In some experiments the pulse is launched from the side at a second sampling pad.

3.2.2. Electrical connections

One strip of the transmission line is held at a fixed voltage with a battery. The second line, above which the STM tip hovers, is held at the (different) STM bias. A pump laser pulse at the generating switch transiently short circuits the two strips and creates a voltage pulse which propagates along the transmission line towards the tip. The transmission line, which is not impedance-matched at its endpoints (but can be if necessary), extends for several millimeters in each direction from the switch to prevent severe reflections from entering the temporal window of the measurement.

The tip is connected to the STM preamplifier, and the sampling pad is connected to an ITHACO 564 current preamplifier with a variable gain of up to 10^8 V/A. When the probe pulse is directed to the sampling pad, the current which comes through the ITHACO preamplifier as a function of time delay reflects the time evolution of the pulse on the sample at the sampling switch. When the probe is directed to the switch on the tip, we resolve in time the current which comes through the tip.

3.2.3. Pulse characterization

The solid line in Figure 3.3a is a voltage pulse generated and measured on a sample with the geometry illustrated above. The pulse is 1.1 ps FWHM, is unipolar and exhibits very little ringing. The time response and shape of the signal are reproducible within about 10% from sample to sample. The dotted trace is the result of the same measurement performed on another sample fabricated in a different run than the first, but with identical design and implantation parameters. It has a 1.2 ps FWHM.

These pulses were measured at a distance of 200 μm from the generation point. At much smaller distances, 0 to 50 μm, the pulse often has more structure associated with the excitation of non-propagating modes[2]. Since our generating and sampling pads are fixed, we are unable to study these changes in the pulse shape along the transmission line with on-sample measurements, but we can use the USTM to perform such a measurement.

Figure 3.3b is a measurement of the signal in 3.3a (solid trace) over a larger interval of time delay. We see a series of bumps appear after a time delay of about 45 ps. This signal, which also appears in the time resolved tunneling measurements, corresponds to a reflection from the end of the transmission line. This reflection is consistent with propagation along the 3 millimeter separation between the switch and the unterminated ends of the transmission line at a group velocity of c/2.3.[3]

The short, clean nature of the pulse in Figure 3.3a derives from the effective coupling of the excitation at the generating switch to the quasi-TEM mode of the transmission line. If we excite from a sampling pad on the side of the transmission line, a multimode excitation results. A pulse generated in this manner appears in Figure 3.3c. The pulse, with a 2.2 ps FWHM, is wider and has stronger ringing than the quasi-TEM pulse.

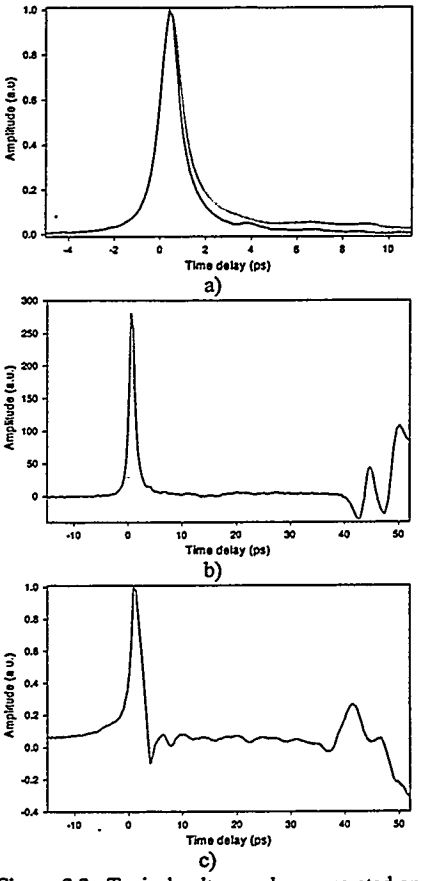


Figure 3.3. Typical voltage pulses generated on our transmission lines. a) Pulses generated with the geometry in Fig 3.2. The solid and dotted signals were generated on samples with identical geometry. b) The solid trace in "a)" measured over a larger time interval. The reflection at ~ 40 ps arises from an impedance mismatch at the end of the line. c) The signal measured when the voltage pulse is introduced from a side sampling pad.

3.2.4. (Non) Invasiveness of the STM

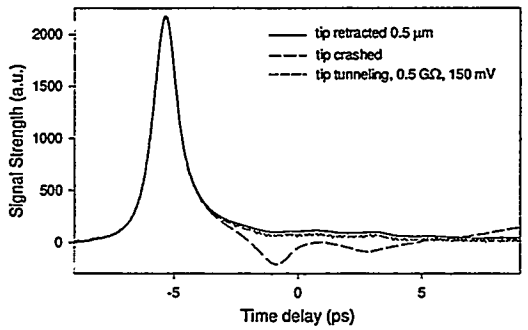


Figure 3.4. Effect of the tip on the sample response. The solid line represents a voltage pulse measured via photoconductive sampling on the sample while the STM tip is 0.5 μm away. The dashed trace with noticeable ringing is the same signal measured while the tip is in ohmic contact with the sample. The short dashed trace is the same signal measured while the tip is in tunneling range.

The sample response is affected very little by the presence of the tip in tunneling range, but is affected somewhat when the tip is crashed. This is illustrated in Figure 3.4. For these data, the sample had the geometry of Figure 3.2, and the tip was placed approximately 150 μm past the sampling switch on the transmission line. Changes in the signal are due to reflections back from the discontinuity in the transmission line at the tip. Similar results also were obtained when the tip was placed between the generating and sampling switches.

The solid trace in Figure 3.4 is the signal measured while the tip is approximately 0.5 microns above the plane of the sample. This trace is identical to the signal which is measured when the tip is not over the sample at all. The dashed trace with several small rings after the main peak is a measurement of the pulse on the sample while the tip is in ohmic contact with one line of the transmission line. Presumably, the ringing results from voltage pulses which reflect at the discontinuity on the transmission line

where the tip makes contact. The ringing produces a peak in the Fast Fourier Transform (FFT) spectrum at 320 GHz, corresponding to round-trip propagation along 200 μm at a velocity of c/2.4. The ringing also may be caused by changes in the pulse propagation due to the stub geometry in the transmission line where the tip makes contact. Some work by other groups using contact probes has shown that there is even less distortion of the signal when the tip is very short[4].

The short-dashed line is the same measurement performed while the tip is tunneling at 0.5 GΩ resistance on the order of 5 Å from the sample. The trace is virtually indistinguishable from the undistorted measurement. Our tunneling measurements are electrically completely non-invasive.

3.3. Experimental procedure

3.3.1. Time calibration

The crashed measurement facilitates calibration of the position signal from the fast scanner. First, we use the fast scanner to measure the crashed signal. Then, we use the Klinger stepper motor to change the time delay by a known amount and use the fast scanner to re-measure the signal, which is now shifted in time. Relating the separation of the peaks, plotted as a function of the position signal, to the known delay between them gives us the required conversion between position signal and time delay. The position signal from the galvanometer is nearly linear with true position, so we only need a few traces at different time delays to get a very accurate interpolation.[5]

3.3.2. Tip approach and laser alignment

During the approach of the tip to the sample, it is impossible to keep the probe beam aligned to and focused on the switch on the tip. To avoid crashing, we set the voltage bias and current set point such that the desired impedance is several times that of the off resistance of the switch. Typical values for loop parameters are a 0.5 nA set-point and 500 mV bias. These values give an impedance of 1 GΩ, well above the 100 MΩ or so impedance of the switch.

With both the laser beams blocked, the STM tip is brought into tunneling range. To align the probe beam to the tip, the tip is brought into a soft ohmic contact with the surface using the Z motion of the piezos. We unblock the probe beam and adjust its position on the switch to maximize the average DC current which flows to the preamplifier. This current results from the DC bias of the STM rather than from a short voltage pulse. We further optimize the alignment by unblocking the pump pulse, scanning the time delay, and maximizing the amplitude of the resulting crashed signal.

After alignment, the average impedance of the switch is reduced to a few MΩ or less. The bias is returned to normal levels of 10-100 mV, the set point is adjusted to a current on the order of 1 nA, and we can begin to make time-resolved tunneling measurements. The DC resistance of the tunnel junction while the tip is in tunneling range is typically between 100 MΩ and 2 GΩ, so the impedance of the switch perturbs the low frequency response of the STM by a few percent at most.

3.3.3. On stable tunneling in air

The tunneling gap of an STM in air is almost never air. There are usually films of water and organic oils, layers and crumbs of metal oxides and carbon soot on the sample surface. These contaminants have several effects on tunneling. First, the barrier height is reduced from the vacuum level to the valence level of the insulating impurity. Second, the dielectric constant in the gap is increased from the vacuum level, so image charge reduction of the barrier height is reduced[6]. These two effects oppose and at least partially cancel each other. The net result is that the exponential decay length for the tunneling current is increased from the value of around 0.5 Å found in vacuum tunneling but is still on the order of 1 Å.

Contaminants may interact mechanically with the tip as well. Several authors have suggested that the tip and/or sample deform as the tip compresses surface contaminants[7]. Hence, a voltage on the piezos which normally would produce vertical motion of N Angstroms reduces the true tip-sample separation by less than N Angstroms. This effect can produce anomalously large corrugations in atomic images on soft materials like HOPG. It also gives rise to an apparent decay length on the order of 10 Å[7]. The true decay length is still on the order of 1 Å, but it may take an applied voltage expected to produce 10 Å of piezo motion to produce this 1 Å change in the gap width due to compression of the barrier contaminants and deformation of the tip.

Before attempting time resolved tunneling experiments, it is important to verify that the STM is tunneling and is functioning normally at DC. We make sure that we can change the current smoothly by changing the current set point in the feedback electronics. This ensures that the tip is not wedged into an incompressible piece of dirt. We

make sure the tunneling I-V curve (on metals and graphite) is nearly linear, and we measure I-Z curves and make sure they are reproducible and have the expected exponential dependence.

These tests verify that the current arises from tunneling through the barrier we expect, rather than from tunneling through mid-gap states or from other channels. The most important test is to ensure that the STM can make reproducible images of the sample with a resolution on the order of 50 Å or better with the lasers on. This again verifies that the tunneling current has a normal variation with distance, ensures that the lasers are not adding a significant source of noise, and verifies that the electronics are functioning properly. The figures in Chapter 2 illustrate some acceptable images.

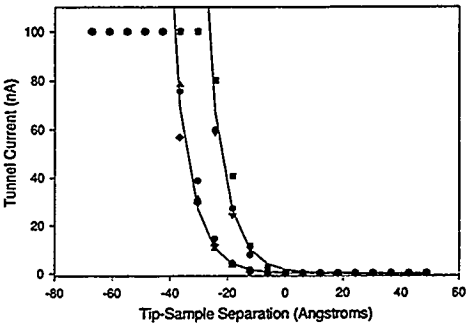


Figure 3.5. Typical I-Z curves for the STM in normal DC operation. The experimental data is represented for three consecutive back and forth scans by circles, squares, and triangles. The solid lines are exponential fits to the curves with 1/e decay lengths of 6.5 Å.

Figure 3.5 illustrates an acceptable set of I-Z curves, measured on a gold transmission line with a Pt-Rh tip using the new microscope. The circles, squares, and triangles each represent a scan from contact to out of range and back. The hysteresis in forward and reverse motion is caused by hysteresis in the motion of the piezos. The flat region in the top left of the graph is the saturation of the preamplifier at 100 nA. To estimate our effective attenuation constant, we fit the decays to the exponential curves represented by solid lines in the figure. The curves have a 1/e decay length of 6.5 Å. This is on the order of 10 Å, as we expect for tunneling in air. The distance scale in the graph is only accurate to within about 20%[8].

3.4. Fast and slow scanning revisited

3.4.1. Terminology

In conventional STM, images can be collected in "current" or "topographic" mode. In current mode, the image displays tunneling current as a function of lateral tip position, and the scan rate of the piezos is set fast enough that the STM feedback loop can not follow changes in the signal. In topographic mode, the scan rate is slow enough that the feedback loop can compensate for changes in the current by adjusting the height of the tip above the surface. The image is formed from the correction voltage sent to the Z piezo electrodes by the feedback loop.

In analogy with the modes of image data in conventional STM, there are also two modes of data collection for single point ultrafast measurements in USTM. In a single point measurement, the time delay can be fast-scanned, so that the feedback loop can not respond, or slow-scanned, so that the tip can move to compensate changes in the signal. In fast-scan mode, one measures the tunnel current, and in slow scan mode, one measures the correction voltage sent to the Z piezo electrodes. The choice between the two does not hinge only on the noise performance discussed in Chapter 2, but also on the effect of the feedback loop on the measurement and on the type of experiment we wish to perform.

If we want to study the properties of the tunnel junction, we require the gap width to remain constant. This excludes slow scans when the signals involved are large enough to cause the gap width to change significantly during the course of the measurement. If we wish to measure processes on the sample, either fast or slow scans should suffice, and the relationship between fast and slow scan signals should be dictated by the measurable DC relationship between the current and voltage and distance. For small changes in the tip to sample bias, this relationship is dominated by the exponential I-Z relationship.

3.4.2. Comparison of Z and I measurements

In the data discussed in this chapter, the current mode (or fast scan) signal consists of a large DC baseline and a much smaller AC rider which contains all the time-resolved information. The topographic mode (or slow scan) signal consists of an AC component which varies with time delay and a DC baseline which can be set arbitrarily. Each signal is measured in volts by an A/D converter. We can write for the measured signals:

$$S_{slow} = S_{slow}^{DC} + S_{slow}^{AC}$$
$$S_{fast} = S_{fast}^{DC} + S_{fast}^{AC}$$

We expect that for tunneling current the signals are related by the standard I-Z tunneling relationship:

$$C\, e^{-A d S_{slow}} = S_{fast}$$

where A is the inverse tunneling attenuation constant discussed above, d is the conversion factor between angstroms

of piezo motion and applied voltage, and C is a geometrical factor which includes the tunneling electrode size[6]. Then,

$$A d S_{slow}^{AC} + constant = \ln\left(1 + \frac{S_{fast}^{AC}}{S_{fast}^{DC}}\right) \approx \frac{S_{fast}^{AC}}{S_{fast}^{DC}}.$$

Hence, the slow scan signal and the fast scan signals are linearly related, up to a DC baseline, with a proportionality constant given by the equation above. This is illustrated for signals measured with different tips and samples in Figure 3.6. The solid line in Figure 3.6a represents a typical fast-scan signal with its baseline subtracted, and the dotted line represents the signal measured in topographic mode. Within the noise of the measurement, the signals are nearly identical. A comparison of the amplitudes of the two signals gives A^{-1} = 5.3 Å. The scans in Figure 3.6b, which have much better time resolution and less ringing, give A^{-1} = 11.4 Angstroms. Each of these inverse decay lengths is within a reasonable range for the apparent inverse decay length of an STM operating in air. The decay length of the time resolved signal measured in true tip-sample separation is therefore on the order of 1 Å, just as in normal DC tunneling.

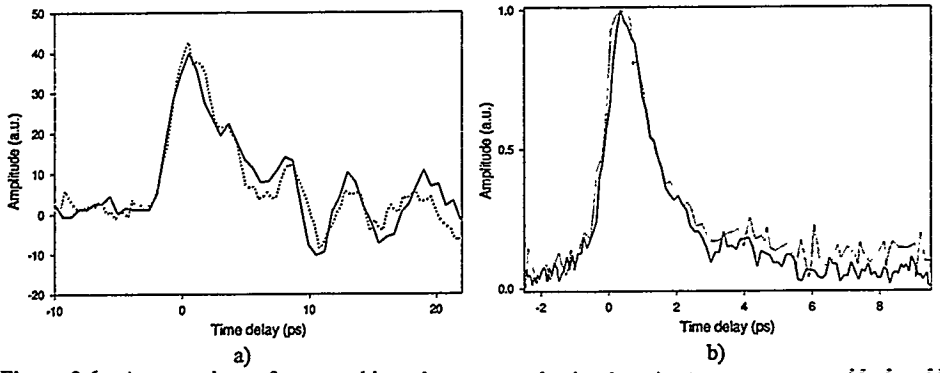


Figure 3.6. A comparison of topographic and current mode signals. a) A measurement with the old microscope with low time resolution. (solid line - tunnel current, dotted line - z) b) A measurement with high time resolution on the new microscope. (solid line - tunnel current, dotted line - z).

Assuming a fairly large A^{-1} of 2 Å and a smooth tip with radius of curvature of 1000 Å, a simple calculation shows that the spatial resolution of the time resolved measurement is approximately 100 Å. In practice, asperities on the end of the tip dramatically improve the spatial resolution. For example, the STM routinely achieves near atomic or atomic resolution in normal operation with a similar attenuation constant. The 100 Å estimate should be considered a conservative lower bound for the resolution of the time-resolved STM.

3.4.3. Feedback effects

Ideally, fast scanning should be done at hundreds of Hz or faster, far above the bandwidth of the feedback electronics. With our present scanners, time-delays of several tens of picoseconds or more can be scanned at a maximum of several tens of Hz. When the bandwidth of the feedback loop is adjusted to be as low as possible, there is a small, residual effect on the signal. At larger feedback bandwidths, the current signal can be significantly distorted.

There are two easy methods to uncover distortion from feedback in the USTM signal. These are illustrated in Figure 3.7. Figure 3.7a shows a signal measured over two complete cycles of the fast scanner. The shape of the signal as a function of time delay over forward and reverse paths is different because the feedback loop sees and responds to different signals on each path. Typical symptoms of feedback include a broad negative dip at negative time delay or a very slow rise time in the main peak.

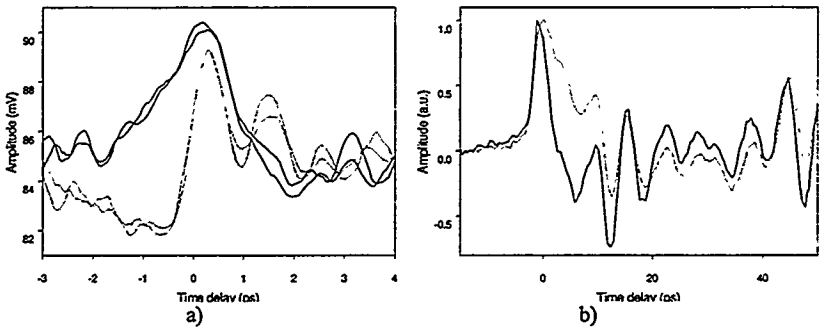


Figure 3.7. The effect of feedback. a) A tunneling signal which is clearly affected by feedback. The signal's shape changes depending on which direction time delay is being scanned. The solid lines represent consecutive scans in one direction, while the dotted lines represent scans in the reverse direction. b) A signal measured by fast scanning at 20 Hz (dotted) compared to the same signal measured at 5 Hz (solid).

Figure 3.7b shows the same signal measured with a 5 Hz scan frequency and with a 20 Hz scan

frequency. As the scan frequency rises above 20 Hz, the signal shape remains the same. If there is a suspicion that feedback effects are present, we increase the scan frequency and decrease the bandwidth of the feedback loop until the signal does not vary with changes in the parameters of the feedback loop or changes in the scan frequency.

Another strategy to eliminate feedback effects is to drive the time delay scanner with a trapezoid shaped signal. The USTM measures current on the way up or on the way down the trapezoid's sides while the feedback is turned off. On the top of the trapezoid, the feedback loop is turned on and the STM is stabilized. If the time spent on the ramps up and down is relatively short and the time on the flat top is relatively long, the STM should run stably, and the signal can be measured only when the feedback is off. We had limited success with this approach, possibly because the baseline noise level of the STM rises too dramatically when feedback is turned off.

3.4.4. Shaker noise

As mentioned in Chapter 2, the vibrations from the shaker couple to the STM. This noise is particularly insidious because it is synchronized with the scanner and therefore is not reduced by averaging many scans. Sometimes this noise primarily affects the baseline of the time resolved signal. In this case, turning the feedback bandwidth up slightly can correct the baseline without distorting the signal.

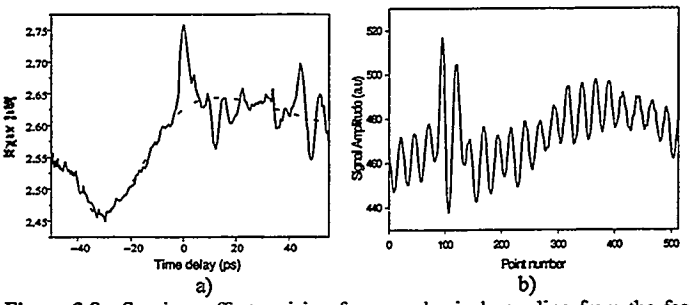


Figure 3.8. Spurious effects arising from mechanical coupling from the fast scanner to the STM. a) A distorted baseline. The signal is indicated by the solid line, and the mechanical background is illustrated with the dashed line. b) High frequency wiggling. The true signal, which has a peak near point number 100, is significantly distorted by vibrations in the STM induced by mechanical wobble in the scanner.

A dramatic example is illustrated in the time-resolved tunneling signal in Figure 3.8a. The distorted baseline, represented by a dashed line, appeared in the signal measured without the pump beam but did not appear in the crashed measurement. These two observations and the fact that the effect was less pronounced when the bandwidth of the feedback loop was increased strongly suggest that the effect is mechanical in nature. Distortions in the baseline from mechanical coupling usually are not so severe, consisting only of a 5% linear rise across the measured signal which can easily be removed.

Another pernicious effect is the high frequency noise shown in Figure 3.8b. This noise occurs as a result of reproducible wobble in the linear translation stage of the scanner. This type of noise is difficult to separate from the true signal. One method to detect it is to look at the variation of the signal with scan amplitude. As the amplitude of the scan is decreased, the position of noise peaks from mechanical coupling does not change, but their amplitude is reduced. The real signal has the opposite behavior - its amplitude remains constant, but its position in the scan changes. In practice, this noise only appears when the scanner has degraded severely and requires repair.

3.5. Features of time resolved signals

The next several sections describe the primary features of time resolved tunneling measurements. These features and their dependencies on voltage and tip sample distance have been inferred from observations of many signals generated from differently shaped voltage pulses on a variety of samples and measured with a variety of tips. The earliest signals were measured with long brass tips, while later signals were measured with short and long Pt-Rh tips and short Ag epoxy tips. All the data shown below was measured in fast-scan mode. None of the data has been smoothed, but in some cases a background has been subtracted. Since much of the data was taken with very short averaging times and under different conditions, the signal to noise levels will vary among sets of data. Except where noted below, we have excluded any data which might be corrupted through feedback effects or mechanical coupling.

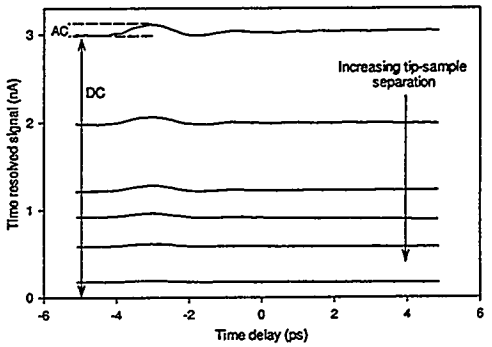


Figure 3.9. AC and DC parts of time resolved tunneling. A series of traces are shown. The time-resolved measurement of tunneling current contains an AC component which carries temporal information and a DC baseline determined by the setpoint of the STM feedback loop.

3.5.1. AC and DC components

The time-resolved current signal measured by the STM in fast-scan mode consists of a small time varying, or

AC, component which rides on a large DC background. The AC component arises from the voltage pulse on the sample and contains all the temporal information in the measurement. The background is simply the nominal DC current used for distance feedback. This current is determined by the set point of the feedback loop and is equal to the bias voltage divided by the sum of the gap resistance and the DC average switch resistance. The two components, AC and DC, are illustrated for a set of time resolved traces measured for increasing tip-sample separation in Figure 3.9. The amplitude of the AC component is typically on the order of 1% of the DC baseline. Both AC and DC components decrease rapidly with increasing separation. In fact, we will find that the distance dependence of both components is identical.

3.5.2. Time resolution

Figure 3.10 illustrates two tunneling signals measured with two tips of dramatically different length. In each measurement, the tip was placed approximately 300 μm from the generating switch on the gold transmission line. In Figure 3.10a, the generating switch connected the main line to a side sampling line biased to 27 V. The signal was detected by the STM while tunneling at 5 nA with a +80 mV bias on the transmission line strip. The tip was a brass wire approximately 500 μm long. The solid trace in Figure 3.10a is the tunneling signal from Figure 3.8a corrected for the very large shaker-induced background. Although the distorted background may make some quantitative statements about the signal suspect, the data is presented here for its historical significance - it is one of the first measurements we performed with the USTM. Of course it is also qualitatively and quantitatively consistent with more recent (undistorted) measurements.

The main peak is 4 ps wide (FWHM) and its 10% to 90% rise time is approximately 2 ps. The dashed trace represents a measurement of the voltage pulse on the sample made independently of the tip. The pulse on the sample has a significantly shorter FWHM of 2.2 ps. Much of the broadening in the tunneling signal relative to the pulse on the sample is a result of dispersion on the tip. This is evident from a comparison of the crashed signal, represented by the dotted trace, to the pulse measured on the sample. The amplitude of the crashed signal is approximately a factor of three larger than the tunneling signal. The crashed signal has broadened from 2.2 ps to 8.5 ps.

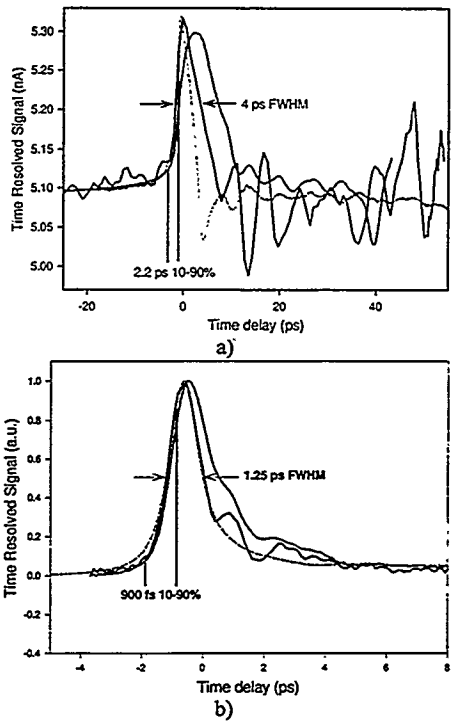


Figure 3.10. Time-resolved tunneling signals (solid) compared to crashed (dotted) and on-sample measurements (dashed) of picosecond voltage pulses with: a) a 700 mm brass tip, and b) a 10 mm silver epoxy tip.

Figure 3.10b shows crashed and tunneling measurements performed with a much shorter tip. In the experiment corresponding to Figure 3.10b, the geometry of the sample was identical to that shown in Figure 3.2, and the bias across the switch was 9 volts. The tip was a sharp cone of silver epoxy on the order of 10 microns long. The dashed curve represents a separate measurement of the pulse on the sample, and the dotted curve represents the crashed signal, which has broadened from 1.2 ps to 1.7 ps. The solid curve is the measured tunneling signal, from which very small linear background has been subtracted. The rise time of the tunneling signal is 900 fs, and the FWHM is 1.2 ps. Except for some slight ringing and a slightly faster rise time, the tunneling signal appears identical to the signal measured on the sample.

3.5.3. Sensitivity

The time resolved signal in 10a consists of a 5% AC signal riding on a 5.1 nA DC background. The height of the peak average tunnel current is increased by approximately 250 pA when the transmission line pulse passes beneath the tip at zero time delay, corresponding to approximately 10 extra tunneling electrons per pulse. The noise level, which can be estimated from the signal at negative time delays, is approximately 10 pA, so the measurement is sensitive to less than one electron per pulse. The amplitude of the pulse on the sample is on the order of 100 mV. The data is an average of 1000 scans at 14 Hz,. corresponding to about 70 seconds of averaging. Therefore the measurement has a sensitivity on the order of 35 $mV/Hz^{1/2}$.

The measurement in Figure 3.10b has higher sensitivity, on the order of 15 $mV/Hz^{1/2}$. The peak height is 60 pA and the noise level is approximately 2 pA with similar integration time. The amplitude of the signal is approximately 6% of the DC background. The additional noise reduction came from faster scanning and averaging and from more

optimal focusing of the probe beam on the switch. The voltage sensitivity can be improved slightly further by bringing the tip closer to the sample to increase the signal amplitude.

3.5.4. Ringing

The tunneling signal in Figure 3.10a exhibits some ringing during the first 20 ps after its main peak and then has a large ring and distortion which occurs approximately 40 ps after the main peak. The signal at 40 ps is the reflection from the end of the transmission line shown in Figure 3.3c. The ringing during the first 20 ps is also clearly exhibited in the voltage pulse on the sample and the crashed measurement.

A glance at the FFTs of these signals in Figure 3.11 shows that while most of the frequencies present in the tunneling signal are also present in the voltage pulse on the sample, there is also a small component at 220 GHz not apparently present in the original pulse. The spectrum is calculated over a window from -40 ps to 30 ps which excludes the large reflection at 40 ps.

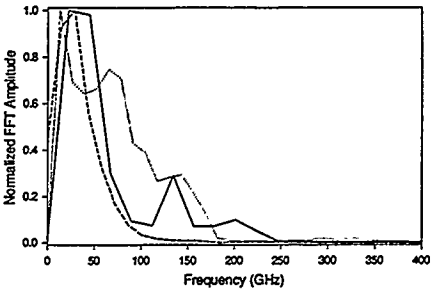


Figure 3.11. The FFT spectra of the data in Fig. 3.10a and of the crashed signal corresponding to this data. The heights of the low frequency peaks have been normalized to 1. The dotted line corresponds to the voltage pulse; the solid line to the tunnel signal; and the dashed line to the crashed signal. The resolution is 13 GHz.

The 220 GHz component probably is ringing which corresponds to reflections which travel back and forth along the tip. The frequency corresponds to a round-trip journey at the speed of light on a 700 μm long tip. This length is reasonably close to our estimated tip length. The amplitude of the component is small, suggesting either that the reflection at the ends of the tip is small or that much of the signal is radiated away as it propagates on the tip[9].

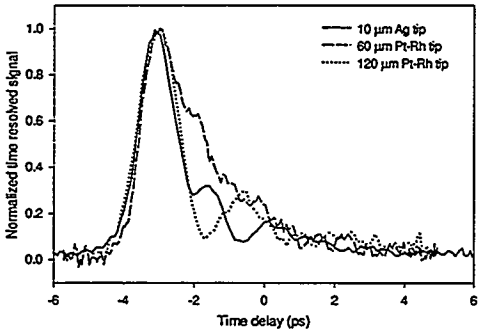


Figure 3.12. Changes in the signal as a function of tip length. A 20% linear baseline was removed from the 60 μm and 120 μm tip signals

In shorter tips, we do not see ringing which can be associated clearly with propagation at *c* along twice the length of the tip. Figure 3.12 shows three measurements of a pulse on the sample under nearly identical conditions but with different length tips. The 10 μm and 120 μm signals have distinct rings with periods of 1.3 and 2.6 ps, respectively. The 60 μm signal appears slightly broader and has less distinct ringing. The ring does not appear in the crashed signals, which merely broaden with increased tip length. The source of the ringing can not conclusively be identified at present.

3.5.5. Linearity

The time resolved tunneling signal is linear over the range of voltage pulses we have been able to generate on the sample, a few hundred mV and below. The traces in Figure 3.13a are a series of time resolved tunneling signals measured for different bias voltages on the generating switch. The generating switch is known to be linear over this range. The pulse is launched from a side sampling line and detected with the STM several hundred μm past the generating switch.

The amplitude of the signals, as measured by the height above baseline of the initial peak, is linear with bias voltage. This linearity is illustrated in Figure 3.13b, which plots the amplitude of the peaks in 3.13a versus bias

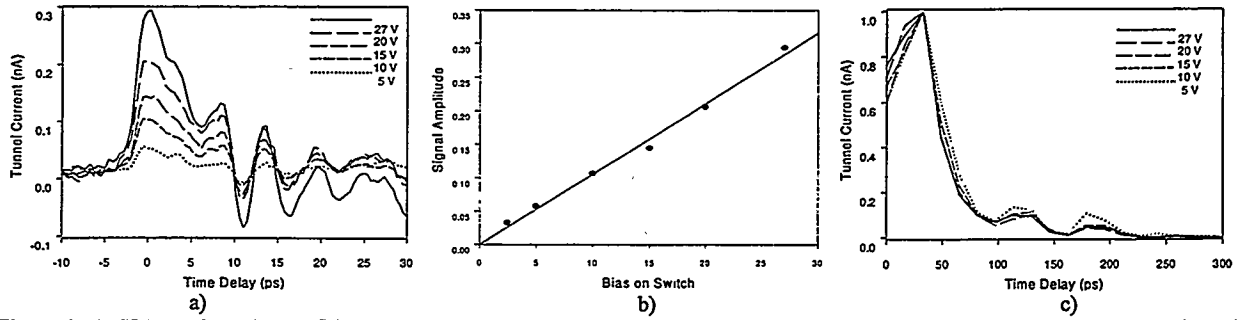


Figure 3.13. Voltage dependence of the tunneling signal. a) A set of traces measured for different bias voltages on the generating switch. b) The amplitude of the measured signals plotted versus bias voltage. c) FFT spectra of the signals calculated over a window from -40 to 30 ps.

voltage. The line is a linear regression forced to go through the origin.

Figure 3.13c shows the FFT spectra of the traces in 3.13a with the low frequency peaks normalized to unity. No new frequency components are generated as bias voltage increases. The slight deviation of the FFT spectra at small bias in the higher frequency components is most likely due to noise. The signal to noise ratio in the measurement at 5 volts bias is only 4::1.

3.6. Distance dependence

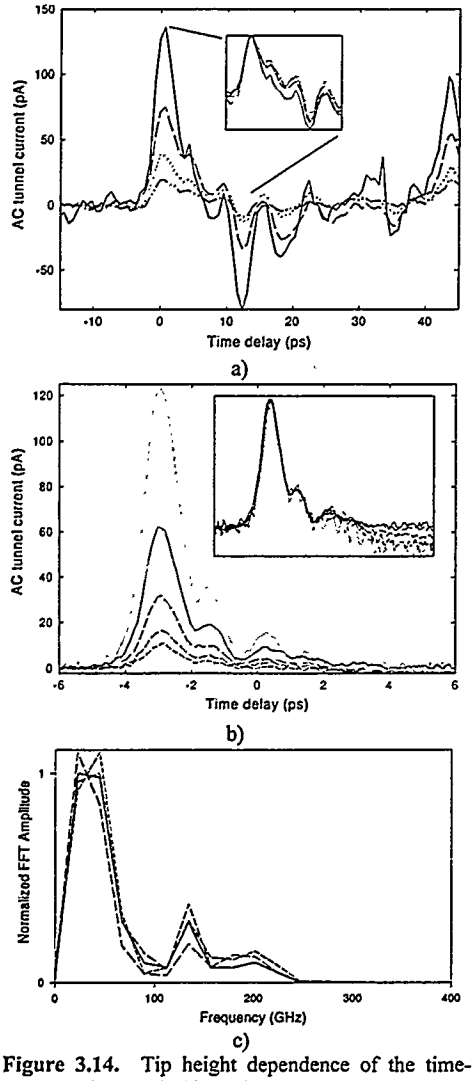


Figure 3.14. Tip height dependence of the time-resolved signal. a) Signal in 3.10a measured for DC tunneling resistances of 32, 64, 128, and 256 MΩ. b) Signal in 3.10b measured for DC tunneling resistances of 75, 150, 300, 600, and 1000 MΩ c) Spectra of the signals in "a)" over a window from -40 to 30 ps.

The distance between tip and sample in STM can be varied smoothly simply by changing the current set point in the feedback loop. The tunneling current in a normal DC measurement decreases exponentially with increased tip height above the sample. As discussed in Section IV, a comparison of fast scanned and slow scanned time resolved signals indicates that the AC component in the time resolved signal also decreases exponentially, with the right decay length, as tip height increases. This dependence is explicitly verified below.

3.6.1. In Tunneling Range

Figures 3.14a and b show a series of time-resolved tunneling traces from which the data in Figures 3.10a and b, respectively, was taken. The signals are measured for a series of different tip-sample separations. The DC background has been subtracted in order to compare the AC parts of the traces. The curves are indexed by the DC tunneling resistance, which is exponentially related to tip-sample separation. In 3.14a, the gap resistance varies from 32 MΩ to 256 MΩ. In 3.14b, the gap resistance varies from 75 MΩ to 1 GΩ.

The insets show the same traces normalized to a peak height of unity. The inset in 3.14a shows only the main peak. It is clear that the shape of the signals, and therefore the frequency dependence of the tunnel junction impedance, does not change significantly with tip height.

Figure 3.14c is a set of FFT spectra of the data in 3.14a. These spectra are calculated for data in a window from -40 ps to 30 ps. This window excludes the reflection around 40 ps and gives 13 GHz resolution.. The heights of the low frequency peaks are normalized, but this normalization is necessarily inexact because the true low frequency peak occurs between data points in the FFT. Nonetheless, it is clear that the signal has a similar distance dependence over a broad range of frequencies.

To study this distance dependence, we can compare the amplitude of the AC component of the time resolved tunneling signal to the DC baseline. Figure 3.15 shows several sets of the AC amplitude plotted versus DC baseline for a representative sample of 5 sets of data, not including the data sets in Figure 3.14, which also exhibit similar behavior. The signals were measured with different length tips and different driving voltages. Sets of identically shaped symbols indicate the data points in each set of measurements, and the solid lines are linear regressions forced to go through the origin.

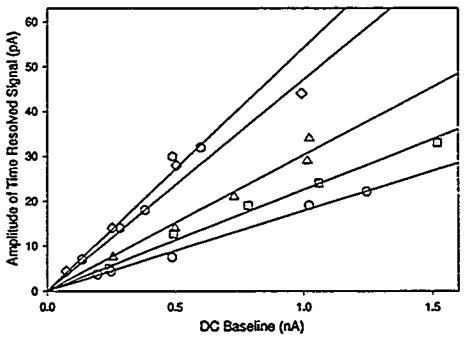


Figure 3.15. AC versus DC. Five sets of data are represented by circles, diamonds, triangles, squares, and pentagons. The solid lines are linear regressions to the data forced to go through the origin. Better linear fits can be obtained with a small AC amplitude at zero DC background.

In all cases, the time resolved signal decreases almost linearly with the DC tunneling signal and therefore has the same decaying exponential dependence for increasing tip sample separation.

Moreover, when the tip is withdrawn from tunneling range, on the order of 10 Å from the surface, the AC component of the signal drops precipitously. In early measurements, the out of range signal dropped below the noise floor and appeared to be zero. Later, more sensitive measurements showed that there is a small but measurable signal out of tunneling range.

3.6.2. Out of range

When the tip is moved out of tunneling range, the amplitude of the time resolved signal drops to a value of approximately 1 pA. This is usually below the noise level for scans on the order of several minutes or less. With averaging times on the order of 30 minutes, it is possible to measure the shape and amplitude of the signal. Since the 1/f noise associated with tunneling is absent in this measurement, better signal to noise can be obtained via lock-in detection.

The distance dependence of the time resolved signal is plotted for microscopic (1 Å) out to macroscopic (10 μm) distances on a logarithmic scale in Figure 3.16. In tunneling range, the distance is calculated from the DC tunneling resistance, assuming a fairly large decay length of 2 Å. Out of range, the distance is calculated from the approximately calibrated piezo motion or is judged from high resolution side-view optical imaging.

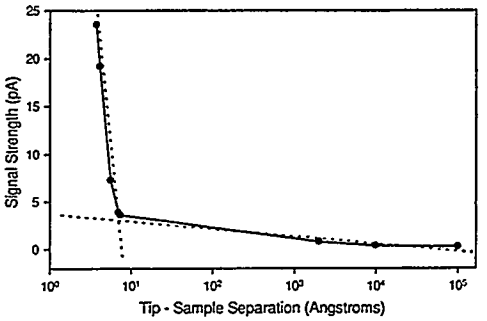


Figure 3.16. Macroscopic distance dependence of the time resolved signal. Note the logarithmic scale on the distance axis.

There are clearly two regimes of distance dependence. In tunneling range, the signal varies exponentially with distance, losing 95% of its amplitude over a range of a few Angstroms. Out of range, the signal remains nearly constant over distances from 100 Å to more than 10 μm. Although there is no data point in the 100 Å range for the set of data in Figure 3.16, we have verified that the signal drops to its long range value immediately after the tip leaves tunneling range. Hence the "corner" in the signal is actually much sharper than it appears.

The shape of the signal also changes when the tip is moved out of range. Figures 3.17a and b show the crashed (dotted), tunneling (solid) and out of range (dashed) signals for two different electrode configurations on the tip. The out of range signals were measured approximately 1000 Å away from the sample surface, and they have been multiplied by a factor of approximately 25. The large rings in the crashed and tunneling signals result from large rings in the voltage pulse on the sample. The signals in 17a and b were measured on different samples. The electrode patterns at the end of the sapphire substrate which holds the tip are illustrated in the top right corner of each figure.

Figure 3.17a illustrates the usual tip geometry - a single electrode connected across the switch to the tip. The out of range signal is more bipolar and has a sharper rise than the tunneling signal. This is indicative of radiative or capacitive coupling. The out of range signal in Figure 3.17b has an unusual shape, and is more distinctly different from its corresponding tunneling signal than is the signal in 3.17a. The data in Figure 3.17b was measured with the lower electrode connected to the STM preamplifier while the upper electrode was floating.

The very slow variation of the AC signal with distance out of tunneling range suggests that the signal is picked up radiatively, either from the tip prior to the switch or from the metal wire after the switch. The fact that the electrode pattern after the switch substantially influences the shape of the out of range signal suggests that most of the pickup occurs after the switch. This is somewhat surprising, because we initially expected to see capacitive coupling to the tip with a stronger dependence on tip to sample separation. Experimentally, we have not observed significant coupling across the geometrical capacitance of the tip in any measurements.

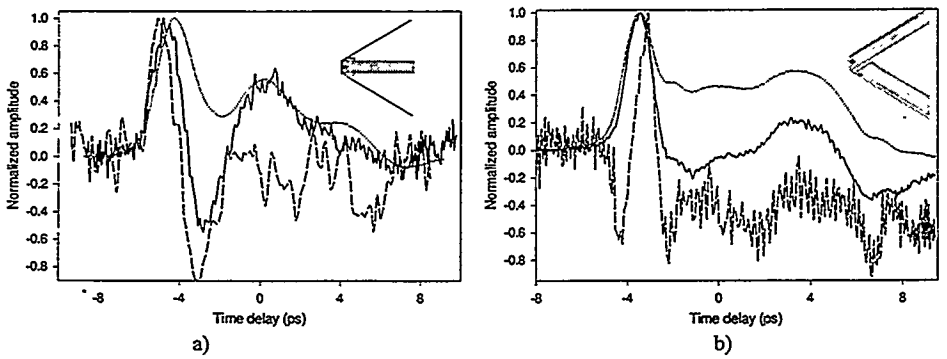


Figure 3.17. A comparison of tunneling (solid) out of range (dashed) and crashed (dotted) signals. a) Signals measured on a "one-line" tip. b) Signals measured on a "two-line" tip. The pattern of electrodes on the tips appears in the top right corner of each figure.

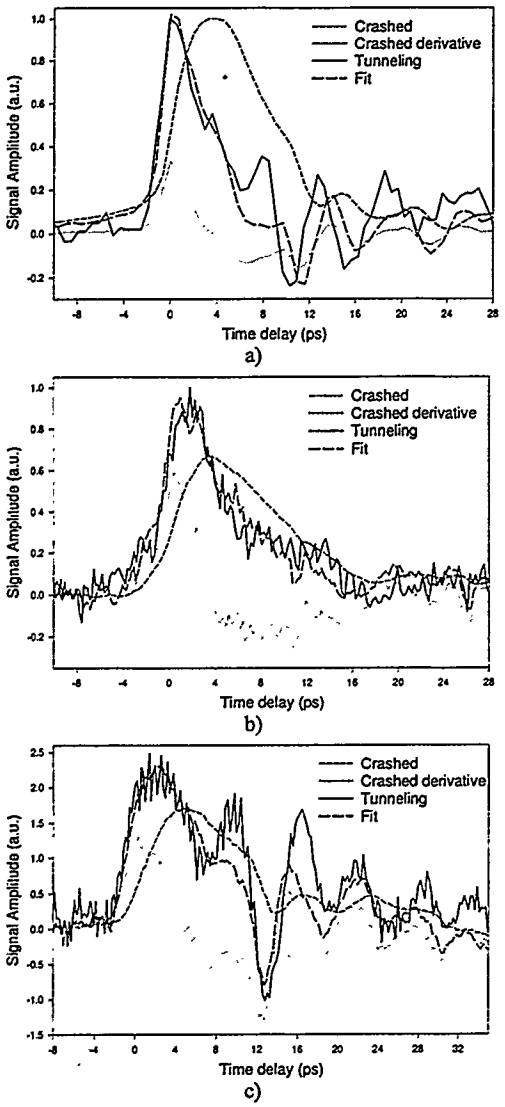


Figure 3.18. Fits to the empirical model $S_{tun}(t) \propto S_{cr}(t) + T\, S_{cr}'(t)$ for fairly long tips. a) Data from Figure 3.10a. The tip was brass, on the order of 500 μm long. RC = 5. b) Tip is 500 μm long Pt-Rh; RC = 3.1. c) Same tip as in "b)", but driven with a different voltage pulse on the sample. RC = 3.3. The fits are consistently good regardless of the shape of the driving signal and the tip material.

3.7. Signal shape - crashed versus tunneling

As we saw in Figure 3.10, the shape of the time resolved tunneling signal differs markedly both from that of the signal measured on the sample and from that of the crashed signal. It is probably more meaningful to compare the tunneling signal to the crashed measurement than to the sample measurement. There are several reasons for this. First, the position of zero time delay changes when we measure with the tip or with the sample, so we cannot quantify shifts in time. Second, the switch on the sample does not have the same geometry as the switch on the tip, so distortions in the pulse shape from the measuring process may be different. Because the voltage pulse on the sample propagates on a transmission line, the different geometry also may change the amplitude of the effective bias across the sample switch and across the junction and tip switch. Finally, we completely ignore the effects of propagation and dispersion on the tip when we compare to the signal measured on the sample. Since both crashed and tunneling signals propagate on the same tip, they should be similarly altered by these effects.

In all our measurements, the tunneling signal has an earlier, sharper rise than the corresponding crashed signal. The main peak is narrower, and any ringing is significantly enhanced. A more quantitative comparison of the crashed and tunneling signals is motivated by the following simple concept. Imagine that the switch samples but does not perturb the "circuit" which includes the voltage pulse, tunnel junction, and switch. It is then easy to show (and it is shown in Chapter 3) that the measured signal should satisfy

$$S_{tun}(\tau) \propto S_{cr}(\tau) + R_t C_t \frac{dS_{cr}(\tau)}{d\tau},$$

where S_{tun} and S_{cr} are the signals measured in tunneling range and crashed, respectively, τ is the time delay between pump and probe, and R_t and C_t are the resistance and capacitance of the tunnel junction, respectively. This simple model might be expected to be valid when the switch is a long way from the junction so that transient signals see the junction without seeing the switch. If we fit the tunneling signal to this model, we can extract R_tC_t. Since we can measure R_t via the DC background, we can extract C_t from such a fit.

In fact, the tunneling signal does seem to fit this model for tips on the order of 500 μm long or longer. Figure 3.18 shows several examples of such fits. In each case, R_tC_t is on the order of 4 ps. A typical value for R_t is 100 MΩ. Therefore C_t must be on the order of 10^{-19} - 10^{-20} F. Furthermore, we know that the shape of the tunneling signal barely changes as we vary R_t over more than an order of magnitude. Therefore R_tC_t remains approximately constant as we vary R_t, and C_t must vary inversely with R_t, exponentially with tip to sample separation!

The size of the extracted capacitance and its exponential variation with distance are somewhat surprising. An estimate for the geometrical capacitance of the junction based on the model of a sphere over a plane gives $C_t \approx 10^{18}$ F only when the tip has an extremely sharp 100 Å radius of curvature. Although it is likely that our tips have asperities this size or smaller, the relatively slow (1/r) variation of electric field with distance means that the geometrical capacitance is less sensitive to such asperities than is the tunneling resistance. It is also physically impossible for a geometrical capacitance to vary exponentially with distance, since this variation requires an electric field whose amplitude increases with increasing distance or electrodes whose size increases with increasing distance. Figure 3.19 illustrates the sharp contrast between the capacitances estimated from our simple model and the geometrical

capacitances associated with very sharp STM tips.

Although the concept of a non-perturbative switch may hold when switch and tunnel junction are far apart, it should not hold when the switch and the STM junction are close enough together to be considered a single lumped circuit. In such a case, when the switch turns on, it not only changes the fraction of the bias from the pulse on the sample which is dropped across the junction, but it also dramatically changes the RC time constant of the circuit containing pulse, switch, and junction. In this case it is not at all clear that a simple relationship will exist between crashed and tunneling signals.

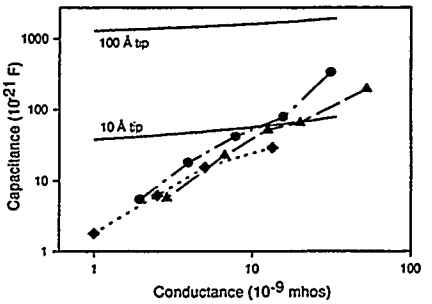


Figure 3.19. Capacitance extracted from fits to the three data sets represented in Figure 3.18. The data, plotted on a log-log scale, has a slope approximately equal to 1, indicating a linear relation between capacitance and conductance. For comparison, the solid lines show the expected geometrical capacitance for extremely sharp STM tips. The extracted capacitances clearly cannot be associated with geometrical capacitance.

Figure 3.20 shows attempted fits when the tip is fairly short, 10 and 120 μm long. The qualitative differences between the two signals are still evident; the tunneling signal has a sharper and earlier rise and a narrower peak compared to the crashed signal. However, there does not seem to be a systematic quantitative relationship between these crashed and tunneling signals. A good fit to the simple model above can be obtained in Figure 3.20a with a value for RC on the order of 0.3 ps. However, the fit in Figure 3.20b requires RC = 3 ps and misses negative dip in the main peak.

There are at least two additional effects at play when we compare fits on long tips to fits on shorter tips. The first is that the signals from long tips have dispersed more, so one might expect the amplitude of the derivative-like component to be increased by a broadening factor. A second change is that for short tips the width of the measured signal is comparable to the width of the switch response. This can lead to an apparent change in the relation between crashed and tunneling signals which might be rectified if the switch response time were much shorter.

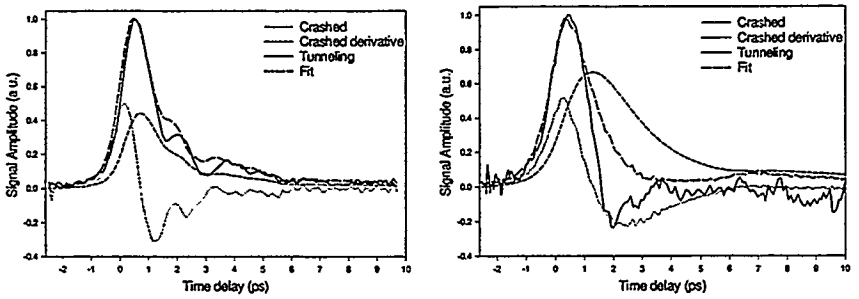


Figure 3.20. Fits to the empirical model $S_{tun}(t) \propto S_{cr}(t) + T\, S_{cr}'(t)$ for short tips. a) 10 μm Ag epoxy tip. b) 120 μm Pt-Rh tip. The fitting parameter now varies with tip length and the fits are not consistently good.

3.8. Antenna effects

3.8.1. Introduction

The time resolved signal in tunneling range has the same dependence as the DC tunneling current on tip sample separation and therefore DC tunneling resistance. There are two ways in which the time resolved signal might depend on this tunneling resistance - either the signal is a result of radiative coupling directly to the tip, which behaves as an antenna terminated by the tunneling resistance, or the signal couples to the tip through the highly local tunneling resistance. If the latter case describes our system, the time resolved signal has spatial resolution similar to that of the STM in normal operation, and we can use the instrument to make time and space resolved images. However, if the the antenna description is correct, the signal should not be expected to have high spatial resolution.

3.8.2. Absence of antenna effects

Several observations indicate that the signal does not result from radiative coupling. Firstly, the tunneling resistance, on the order of 100 MΩ, which forms the terminating impedance of the antenna, is in parallel with the radiation resistance of the antenna, which is on the order of 400 Ω. Changes in the terminating impedance therefore should change the amplitude of the signal coupled to the switch at the other end of the antenna by on the order of 10^{-6}. We have observed instead a linear dependence of the signal on tunneling resistance. Secondly, the strength of the signal does not exhibit variations with the length of the tip, in complete contrast to the result one expects for antenna coupling. As we have seen, the length of the tips varied over a range from much smaller than the wavelengths in the signal out to more than a wavelength of the high frequency components in the signal. Groenveld and coworkers have

recently verified the complete absence of signal variations with tip length[10].

Finally, we have studied the dependence of the signal on the orientation of the tip relative to the transmission line and have not observed a change in the amplitude or shape of the measured signal. The results of such an experiment are shown in Figure 3.21. The top graph in Figure 3.21a shows the tunneling signal measured for orientations of the tip assembly perpendicular and at an acute angle of approximately 30 degrees relative to the transmission line on the sample[11]. The tip, approximately 700 μm long, was oriented approximately 45 degrees from vertical in these measurements. Figure 3.21b shows the crashed measurements corresponding to the signals in 3.21a. The differences between the tunneling signals correspond to changes in the signal generated on the sample as indicated by the differences between the two crashed signals.

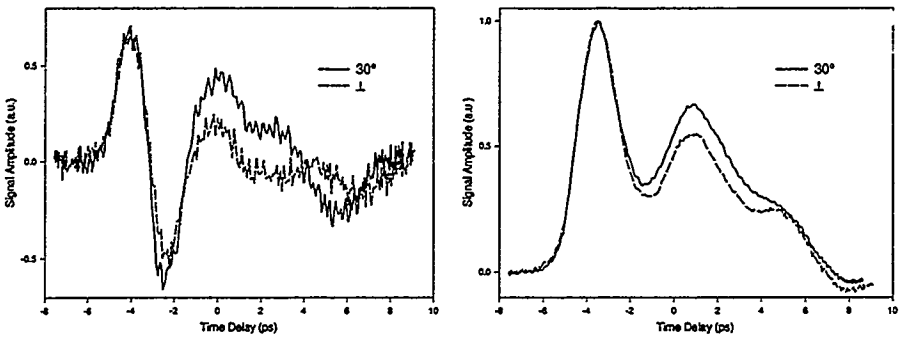


Figure 3.21. Signal dependence on tip orientation. The three traces in each graph are a) tunneling signals and b) corresponding crashed signals measured for orientations parallel (solid trace), perpendicular (lower dashed trace) and at 30 degrees from parallel (upper dashed trace). The slight variations mostly can be attributed to the variations in the shape of the pulses generated on the sample.

3.8.3. Two-line tips and terahertz pickup

It is possible to design a tip which picks up purely radiative signals. We designed the "two line" tip schematically illustrated in the lower right corner of Figure 3.22a to permit alignment to the switch on the tip without crashing. This design also allows us to perform the USTM measurement in a non-perturbative manner closely akin to the sampling measurement on the sample. In practice, however, this tip functioned as an antenna. Although the crashed signal looked identical to the crashed signal measured with the usual "one line" tip, the tunneling signal looked like the one shown in Figure 3.22a.. Figure 3.22b plots the signal amplitude versus the DC baseline and compares it to the dependence seen with a "one line" tip. The signal from the "two line" tip is almost purely radiative and is nearly constant with increasing tip sample separation. The tip used in the measurement of the signals in 17b is an improved version of this tip, with fewer sharp corners. Although it picks up an unusual signal shape out of range, it functions normally while in tunneling range.

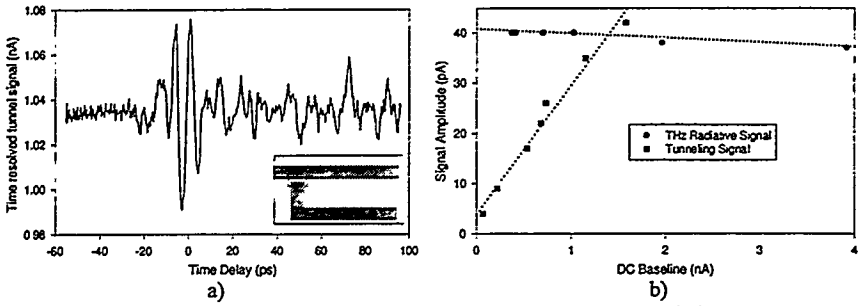


Figure 3.22. Unwanted radiative pickup. a) The signal associated with radiative pickup, and the electrode pattern on the tip assembly. b) The dependence of the radiative signal with distance, compared to a normal tunneling signal.

3.9. Summary

This chapter has examined the experimental features of time resolved tunneling measurements. The tunneling signal consists of a DC baseline and a time varying component. When the voltage pulse on the sample has approximately the same amplitude as the DC bias voltage, the time-varying signal has an amplitude on the order of 5% of the background DC tunneling current. The crashed signal has an amplitude on the order of 100 times that of the time-resolved tunneling signal measured at 1 GΩ.

The most important observation is that there are two distinct regimes delineated by the distance of the tip from the sample. In tunneling range, the amplitude of the signal decays linearly with the tunneling resistance and does not arise from radiative coupling to the tip. Out of range, there is a residual signal reduced by a factor of between 25 and 50 from the amplitude measured at a tunneling resistance of 100 MΩ. The residual component varies slowly over macroscopic distances and is probably due to capacitive or radiative coupling to the tip assembly. The existence of these separate regimes is compelling evidence that the signals in tunneling range are intimately tied to the process of

tunneling to the tip. This in turn suggests that the time-resolved signal can be measured with spatial resolution comparable to that of the STM in normal operation.

While the amplitude of the signal changes as we increase the tunneling resistance, its shape does not. The shape of the tunneling signal is best compared to the crashed signal measured with the same tip. We find that the tunneling signal has a sharper and earlier rise and a narrower peak than the crashed signal. For fairly long tips, we find that the shape of the tunneling signal fits well to the empirical relation $S_{tun}(\tau) = S_{cr}(\tau) + T\,S_{cr}'(\tau)$, where T is on the order of 3-5 ps and remains approximately constant as the tunneling resistance changes. The relationship becomes more complicated for shorter tips and correspondingly faster signals.

Chapter 3 References:

1. M.B. Ketchen, D. Grischkowsky, T.C. Chen, C-C. Chi, I.N. Duling, III, N.J. Halas, J-M. Halbout, J.A. Kash, and G.P. Li, "Generation of subpicosecond electrical pulses on coplanar transmission lines", *Appl. Phys. Lett.*, **48**, p. 752 (1986).
2. Daniel R. Grischkowsky, *private communication.*
3. The expected range for group velocity is between c/2.3 and c/2.5. The uncertainty results from uncertainty in the orientation of the crystalline substrate relative to the transmission line.
See Daniel R. Grischkowsky, Mark B. Ketchen, C-C. Chi, Irl N. Duling, III, Naomi J. Halas, Jean-Marc Halbout, and Paul G. May "Capacitance Free Generation and Detection of Subpicosecond Electrical Pulses on Coplanar Transmission Lines", *IEEE J. Quant. Elec.*, 24, p. 221 (1988).
4. T. Pfeifer, H.- M. Heiliger, E. Stein von Kamienski; H.G. Roskos, and H. Kurz, "Fabrication and characterization of freely positionable silicon-on-sapphire photoconductive probes", *J. Opt. Soc. Amer. B*, **11**, p. 2547 (1994).
5. Unless the gain on the position signal is deliberately changed, we only need to redo the calibration every few weeks to ensure that the performance of the scanner has not deteriorated.
6. E. Clayton Teague, "Room Temperature Gold-Vacuum-Gold Tunneling Experiments", *J. Res. Nat. Bur. Stand.*, **91**, p. 171 (1986).
7. J. H. Coombs and J. B. Pethica, , "Properties of vacuum tunneling currents: anomalous barrier heights", *IBM J. Res. Develop.*, **30**, p. 455 (1986).
H.J. Mamin, E. Ganz, D.W. Abraham, R.E. Thomson, and J. Clarke, "Contamination-mediated deformation of graphite by the scanning tunneling microscope", *Phys. Rev. B.*, 34, p 9015 (1986).
8. We estimate the relation between piezo motion and voltage by comparison with and extrapolation from calibrated piezo tubes with similar geometries. The error may be much less than 20%. When necessary, very accurate calibration can be achieved. The most common way to perform a calibration is to image atomic steps on a crystalline gold surface.
9. A simple estimate from transmission line theory would predict a nearly 100 % reflection at both the tunneling junction and the switch. In reality, the problem involves solving the time dependent wave equation at a complicated discontinuity; and most of the signal at the top of the tip likely radiates into space rather than reflecting. Steven Schwarz, *private communication.*
10. Rogier Groenveld, *private communication*
11. There is an additional effect in the parallel configuration which makes it misleading to compare the parallel signal's amplitude to the amplitude in other orientations. In the parallel case, the tip and the electrodes on the sample may together form a transmission line which changes the propagation of the signal along the tip. In other orientations this effect should be much smaller. The parallel signal shows a 10-20% enhancement in the height of its main peak, in contrast to the expected result for a signal arising from antenna pick-up. The parallel signal should be smaller than the perpendicular signal because the field lines on the sample are oriented along the width of the tip rather than along its length.

Chapter 4
THEORY OF THE ULTRAFAST SCANNING TUNNELING MICROSCOPE

4.1. Introduction

The previous chapter introduced and discussed many significant characteristics of USTM measurements of a voltage pulse. One observation is that when the amplitude of the pulse on the sample is on the same order as the DC bias, the tunneling signal has an amplitude on the order of a few percent of the DC tunneling background. A second is that the strength of the signal varies exponentially with tip sample separation while the tip is in tunneling range. A third is that the signal strength when the tip is out of range is very small. Additionally, the shape of the signal remains essentially unchanged with tip sample separation while the tip is in range. Finally, the main peak in the signal is narrower and has a sharper and earlier rise time than the corresponding crashed measurement. In long tips, the shape of the tunneling signal is empirically described by a linear combination of the crashed signal and its derivative with respect to time delay. Our goal in this chapter will be to understand these results in the context of a simple, physically realistic theory based on a circuit model of the USTM.

The USTM is an extremely complicated electrodynamic system. In principle, to determine the average current to the preamplifier, we must solve the time-dependent Maxwell equations with complicated boundary conditions. As is well known, these equations reduce to Kirchoff's laws when we deal with isolated lumped circuits or with lumped circuits connected to perfect transmission lines. However, it is not entirely clear that a circuit model including these components can describe the operation of the USTM. It is also not clear that the magnitudes of circuit components, particularly capacitances, will have the values we would expect for low frequency measurements. Even in apparently simple circuits containing photoconductive switches, the effective capacitance of the switches can be far different than the expected static geometrical capacitance[1].

A block diagram which illustrates the components of the USTM that might go into a circuit model is shown in Figure 4.1. There are several elements which can lead to difficulties in the model. A long STM tip may act like a complicated nonlinear transmission line or an antenna fed through the tunnel junction, so radiative losses from the tip and dispersion on the tip can have a large impact on the shape and amplitude of the measured signal. Also, the high local current density at the tunnel junction profoundly affects the field distribution on the tip and sample electrodes.

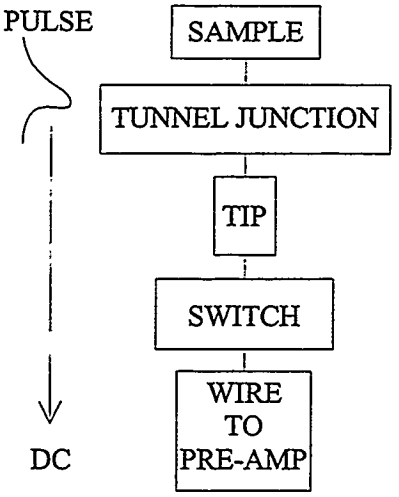


Figure 4.1. The elements of the USTM. A transient pulse propagates through the STM junction, along the tip, through the switch, to the electrode connected to the pre-amp. At that point, only the DC component of the signal is relevant.

If the STM junction and the switch junction are close enough, the two junctions can be described as discrete components in a lumped circuit. The distance scale is set by the velocity of signal propagation on the tip multiplied by the inverse of the largest frequency in the excitation on the sample. When the distance from the tunnel junction to the Auston switch is large compared to this quantity, the tunnel junction and switch can be treated as two independent lumped systems.

As we saw in Chapter 1, the tip is a single conductor which is surrounded partially by air and partially by conductive silver epoxy embedded in a dielectric medium. The propagation speed may be reduced from the speed of light by the effective dielectric constant of the medium surrounding the tip. Therefore it is not entirely clear where the crossover from lumped to non-lumped circuit occurs. For a 1 ps pulse propagating at the speed of light, the crossover should occur at a tip length of approximately 1 mm, but it is conceivable that even fairly short tips on the order of 200 μm long behave as if the switch and STM junction are well separated. It is also conceivable that the photoconductive switch, which extends 50 μm laterally, does not behave as a discrete element.

Figure 4.2 below illustrates a set of possible circuit models for the USTM in a measurement of a transient electronic processes on the sample. The electronic process is modeled as a voltage pulse $v(t)$ roughly centered around $t = 0$ which passes under and biases the STM junction. The STM junction is modeled as a parallel combination of resistance and capacitance. The photoconductive switch is also modeled by a parallel combination of fixed capacitance and time varying resistance $R_s(t-\tau)$.[2] Figures 4.2a and 4.2b show the limiting cases in which we ignore the effects of capacitance on the measurement.

There are three main classes of circuits grouped into the columns of Figure 4.2: lumped circuits under the

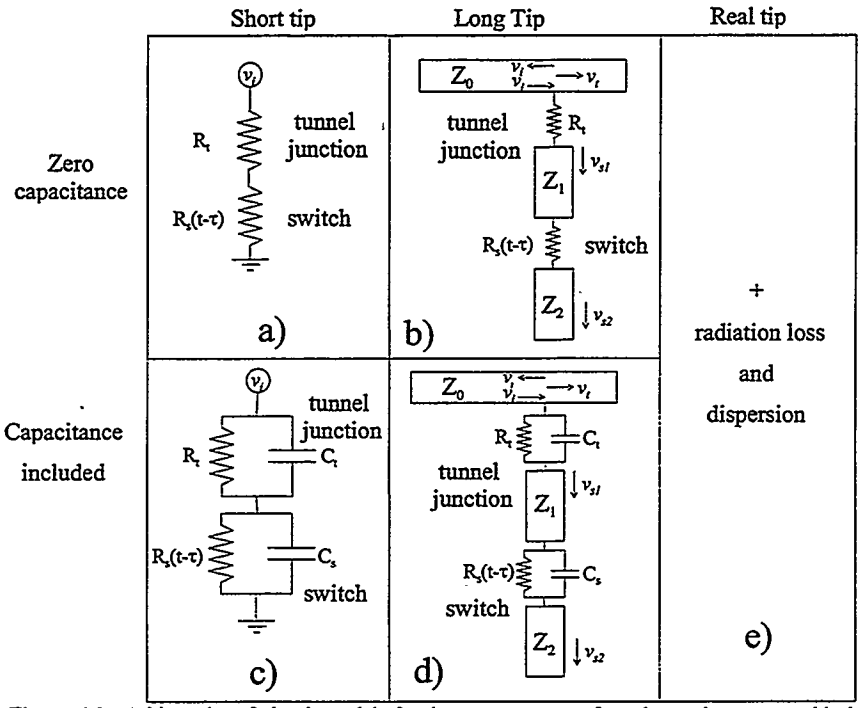


Figure 4.2. A hierarchy of circuit models for the measurement of an electronic process with the USTM a) A "lumped circuit" tip in which capacitive paths are ignored. b) The switch and junction are now separated by a transmission line with impedance Z_1 representing the tip. c) A more sophisticated lumped circuit which accounts for the capacitance of the switch and tunnel junction. d) Separated switch and tunnel junction with capacitances included. e) More accurate models might include the effects of propagation and dispersion on the tip. Ultimately it may be necessary to solve a time-varying wave equation with boundary conditions.

heading "short tip", discrete elements connected by perfect transmission lines under the heading "long tip", and discrete elements connected by lossy, dispersive lines. A thorough treatment of radiation and dispersion from the tip in conjunction with a circuit model requires analytical methods beyond the scope of this thesis[3]. Therefore, I will not analyze the circuits which might appear in the third column. However, I will attempt to mitigate the effects of dispersion on the predictions of the other circuit models by comparing tunneling measurements to crashed measurements performed with the same tip. Because both signals have propagated on the same structure, a direct comparison of the two compensates in part for the effects of dispersion.

This chapter is divided into two main parts. The first part, Section 4.2, will discuss and analyze the circuit models in Figure 4.2. It will compare the predictions of these models to major empirical observations outlined at the beginning of this chapter. As an introduction to the concepts involved in the circuit analysis, the section begins with a brief discussion of Auston switches and simple circuits involving them.

We will find that none of the models in Figure 4.2 is sufficient to explain the entire set of empirical observations. However, some general conclusions still can be drawn concerning the size of the capacitance across the STM junction and the variation of this capacitance with tip-sample separation. The effective capacitance of the STM junction is anomalously small. Further, the capacitance appears to vary nearly exponentially with tip sample separation while in tunneling range. The second part of the chapter, Section 4.3, discusses possible explanations for this behavior and discusses related work by other researchers.

The calculations involved in the solution of the USTM circuit models are straightforward but often elaborate. The reader can safely skip many of the complicated intermediate expressions in the calculations without missing any fundamental details. Equation (17) is the major result of these calculations, and it is used as the basis for the analysis which follows.

4.2. Circuit analysis of the USTM

4.2.1. Auston switches

4.2.1.1. Basic concepts

Ultrafast photoconductive switches, commonly called Auston switches, were studied extensively by Auston and others during the late 1970's and early 1980's[4]. Virtually all analysis of ultrafast measurements with photoconductive switches represents the switches as time-varying circuit elements in lumped circuits which connect one or more transmission lines. We will also use these approximations. As discussed previously, the lumped element approximation is valid only when the dimensions of the elements are small compared to the wavelengths which comprise the signal to be measured. The approximation of perfect transmission lines is valid only when the signal propagates on an electrode with a well-defined ground plane located much closer to the electrode than the largest wavelength in the signal.

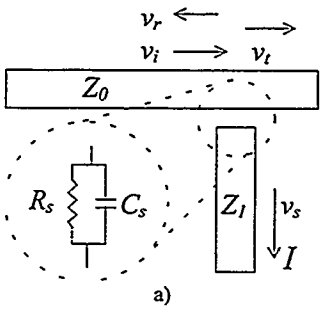
Picosecond photoconductive switches are formed from high mobility, high resistivity semiconductors with

picosecond scale carrier lifetimes. When a short optical pulse illuminates the switch to turn it on, the pulse generates free carriers which dramatically increase the conductance during the carrier lifetime. The free carriers also reduce the capacitance across the switch by reducing the dielectric constant. This change is fairly small[5] and nearly always is neglected. In the high excitation regime, $G_{on} >> G_{off}$, the rise time of the switch response is limited by space charge effects to on the order of G_{on}/C. Typical values for these parameters, as discussed in Chapter 2, are $G_{on} = 100$ Ω and $C = 1$ fF. Hence the rise time cannot be much faster than 100 fs. The fall time is determined by the carrier lifetime and cannot be made much shorter than 500 fs.

4.2.1.2. Simple circuits[6]

Figure 4.3 shows two simple circuits which illustrate common experimental sampling measurements using photoconductive switches. A laser-generated voltage signal $v(t)$ propagates along a transmission line to a junction containing an Auston switch. At the junction, the incident voltage pulse breaks into reflected, transmitted, and sampled components. The switch, which is turned on by a laser pulse at time τ, is modeled as a parallel combination of a time-varying resistance $R_s(t-\tau)$ and a constant capacitance C_s. One uses a current amplifier to measure the average current which flows through the junction.

A straightforward application of Kirchoff's laws to the circuit in Figure 4.3a leads to the equations:



$$v_i + v_r = v_t = \frac{q_s}{C_s} + v_s$$
$$\frac{v_i + v_r}{Z_0} - \frac{v_t}{Z_0} = \frac{v_s}{Z_1} = q_s' + \frac{q_s}{R_s C_s} \qquad (1)$$

where q_s is the charge on the switch capacitor. Solving these equations leads to the following differential equation for the current through the switch:

$$I + T(t)I' = F(t) \qquad (2)$$

where

$$F(t) = \frac{v_i + R_s C_s v_i'}{R_s + R},\ T(t) = \frac{RR_s C_s}{R_s + R},\ \text{and}\ R = Z_1 + \frac{1}{2}Z_0.$$

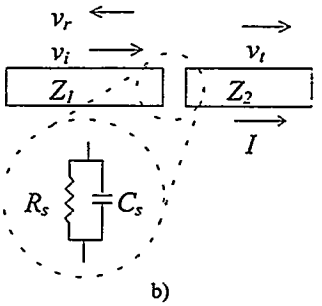


Figure 4.3. Simple junctions containing photoconductive switches. The circuits are drawn as microstrip lines a) A side gap. b) An "in-line" gap.

The circuit in Figure 4.3b satisfies an equation of the same form with

$$F(t) = 2\frac{v_i + R_s C_s v_i'}{R_s + R},\ T(t) = \frac{RR_s C_s}{R_s + R},\ \text{and}\ R = Z_1 + Z_2.$$

The differential equations can be solved explicitly to obtain

$$I = I_0 e^{-\int_{-T_0}^{t} \frac{dt'}{T(t')}} + \int_{-T_0}^{t} F(t') \frac{e^{\int_{t'}^{t} \frac{dt''}{T(t'')}}}{T(t')} dt', \qquad (3)$$

where $2T_0$ is the repetition period of the laser which drives the switch.
The current I depends on the parameters t and τ. The measured signal is given by the time-averaged current

$$\overline{I(\tau)} = \frac{1}{2T_0}\int_{-T_0}^{T_0} I(t,\tau)dt,$$

where $2T_0$ is period over which we average. As discussed in the introduction to the thesis, $2T_0$ usually can be taken to be the repetition rate of the laser driving the switch.
To obtain a simpler result, one usually writes

$$\frac{1}{R_s} = G_s(t) = G_0 + g(t-\tau),$$

where the laser induced enhancement $g(t)$ of the switch conductance is large only in a picosecond wide region around zero. One then expands the terms in the exponents to obtain Auston's result[6]

$$\overline{I(\tau)} \propto \int_{-T_0}^{T_0} v_i(t) f_s(t-\tau)\, dt \; + \text{constant terms},$$

where

$$f_s(t) \propto \int_{t}^{T_0} g(t') e^{-\gamma(t'-t)} dt',$$

and γ is the effective *RC* time constant of the circuit.

From the above expression for the average current, we see that the measurement represents a convolution of the signal to be measured with a sampling function, f_s. The convolution "picks out" the amplitude at time τ of the signal to be measured averaged over the temporal width of the sampling function. The sampling function is proportional to the average laser-induced conductance through the switch from time *t* onward, weighted by an exponential *RC* decay. The width of the sampling function therefore is limited by the width of the switch response and by the *RC* time constant of the circuit.

The above expression for $\overline{I(\tau)}$ is truly elegant, so it is unfortunate that the major approximation involved is not valid in our USTM experiments. The approximation requires

$$\left| \int_{t'}^{t} \frac{g(t)}{C_s} \right| << 1 \text{ for all } t, t'. \qquad (4)$$

Physically, this requirement states that the on-state RC time constant of the switch is much larger than the width of the conductance peak of the switch. In this case, no charge leaks off the capacitor while the switch is on, and the switching action does not perturb the circuit. For a typical C_s of 1 fF and a typical temporal width of *g(t)* of 1 ps, this criterion is satisfied when $g(0) \leq 10^{-4}$ MHos. Despite the fact that this condition is rarely, if ever, satisfied, the simplified results are used freely in the literature to analyze photoconductive sampling measurements. In our experiments, *g(0)* is on the order of 10^{-2} MHos or larger, and no such expansion is possible.

4.2.2. Zero-capacitance USTM models

The circuits illustrated in Figure 4.2a and 4.2b are the simplest models one can envision for the USTM. The tunnel junction is modeled as a resistor with no capacitance, and the switch is modeled as a time-varying resistance with no capacitance. In 2a, the two elements are lumped, and may describe a short tip. In 2b, the switch and tunnel junction are connected by a transmission line with characteristic impedance Z_1.

4.2.2.1. Lumped circuit - zero capacitance

In 2a, the current at any time is given by

$$I(t,\tau) = \frac{v_i(t)}{R_t + R_s(t-\tau)}. \qquad (5)$$

The signal we measure is the time averaged current

$$\overline{I(\tau)} = \frac{1}{2T_0} \int_{-T_0}^{T_0} I(t,\tau) = \frac{1}{2T_0} \int_{-T_0}^{T_0} \frac{v_i(t)}{R_t + R_s(t-\tau)} dt', \qquad (6)$$

where $2T_0$ is the repetition period of our measurement, approximately 13 ns.

In order to gain some more insight into this result, consider the case of a stepwise continuous switch response, in which the switch is on only in the interval $[\tau,\tau+\Delta]$.. The switch has a constant value R_s while off and a constant value $R_{s\text{-}on}$ while on. The time averaged current then becomes

$$\overline{I(\tau)} = \frac{1}{2T_0} \int_{\tau}^{\tau+\Delta} \frac{v_i(t)\cdot(R_s - R_{s-on})}{(R_t + R_{s-on})(R_t + R_s)} dt' + \int_{-T_0}^{T_0} \frac{v_i(t)}{(R_t + R_s)}. \qquad (7)$$

The second term is a DC offset equal to the average current that would be driven through the circuit if the switch were off over the repetition period of the measurement. The first term represents a sample of the current at time τ over a temporal window of width Δ. We saw in Chapters 2 and 3 that reasonable values for the parameters above are

$\Delta \approx 1$ ps, $R_s \approx 100$ MΩ, $R_{s-on} \approx 100$ Ω, and $R_t \approx 100$ MΩ. The crashed signal is obtained by letting R_t become small, on the order of 10 Ω.

The predictions of this model substantially disagree with our empirical observations. First, the calculated signal is proportional to the crashed signal alone. In contrast, we always observe a distinct difference between the shapes of the crashed and tunneling signals. Second, the amplitude of the calculated signal is reduced by a factor of $\Delta/2T_0$ from the amplitude we would measure for a DC signal. This reduction is on the order of 10^{-4}, in contrast to the 10^{-2} reduction we observe, for a 1 ps switch response and our 13 ns repetition period. Third, the ratio of the amplitude of the crashed signal to that of the tunneling signal is on the order of 10^6 assuming a crashed resistance on the order of 100 Ω in the STM junction. This contrasts with our observation that the ratio is on the order of 100 for tunneling measurements at 1 GΩ. Additionally, the variation with distance is not simply proportional to $1/R_t$. However, this last variation would be difficult to detect in the data of Figure 3.15. It would appear only as a slight vertical offset from the origin.

4.2.2.2. Separate junction and switch with connecting transmission line

The circuit in Figure 4.2b can be solved by analogy with the simple Auston switch circuits discussed in the previous section with capacitances set to zero. The voltage pulse is unaware of the presence of the switch as it biases the tunnel junction, so. it generates a pulse on the tip in analogy with the circuit in Figure 4.3a. This pulse propagates to and becomes the incident voltage at the switch. The current through the switch is solved in analogy with Figure 4.3b.

In principle, we could also include contributions from pulses which couple to the tip, reflect at the switch and then the junction to return to the switch. However, we found experimentally in Chapter 3 that the contribution to the signal from this type of reflection is extremely small for tips longer than a few hundred microns. Neglecting the contributions of such doubly reflected pulses, we find that the instantaneous current is given by[7]

$$I(t,\tau) = \frac{2Z_1 \cdot v_i(t)}{\left(R_t + Z_1 + \frac{Z_0}{2}\right)\left(Z_1 + Z_2 + R_s(t-\tau)\right)}. \qquad (8)$$

The time-averaged current is therefore

$$\overline{I(\tau)} \approx \frac{1}{2T_0}\int_{\tau}^{\tau+\Delta} \frac{2Z_1 \cdot v_i(t)}{\left(R_t + Z_1 + \frac{Z_0}{2}\right)\left(Z_1 + Z_2 + R_s(t-\tau)\right)} dt \qquad (9)$$

The transmission line on the sample has a design impedance Z_0 on the order of 100 Ω. The parameters Z_1 and Z_2 may have much higher impedances, possibly on the order of 10^4 Ω or higher, due to the lack of a nearby ground plane. However, the parameters Z_0, Z_1 and Z_2 will almost certainly be small in comparison to R_t and $R_{s\text{-}(off)}$, with magnitudes on the order of 10^8 Ω.

In this model, the signal is roughly proportional to $1/R_t$ as we observe experimentally. However, it is clear that the major problems of the previous model persist. The shape of the calculated signal is again similar to the shape of the crashed signal alone, and the amplitude of the calculated signal is two orders of magnitude too small. It is possible that dispersion accounts for the change in shape that we observe, but the effect should then depend strongly on the geometry and construction of the tip, and the shape of the tunneling signal should not have the simple relationship to the crashed signal that we observe for long tips.

4.2.3. Capacitive models of the USTM

The zero-capacitance models used above are clearly inadequate. Neither are they physically plausible, nor do they correctly predict the shape of the signal we observe. Yet they serve as a motivation for and introduction to the more complex circuits of Figures 4.2c and 4.2d which include the parallel capacitances of the switch and the tunnel junction. In this section I consider the circuits of Figures 4.2c and 4.2d and the limiting cases in which one or the other of the capacitors is very small.

4.2.3.1. Transmission line- Figure 4.2d

The tunnel junction in Figure 4.2e looks exactly like the circuit in Figure 4.3a. Hence we find that the voltage pulse which passes through the tunnel junction satisfies the equation

$$I_t + T_1 I_t' = F_1(t), \tag{10}$$

where

$$T_1 = \frac{R_t R C_t}{R_t + Z_1 + \frac{Z_0}{2}},\ F_1(t) = \frac{v_i(t) + T_t v_i'(t)}{R_t + Z_1 + \frac{Z_0}{2}},\ \text{and } T_t = R_t C_t.$$

This equation describes a current driven by the "effective voltage" term F_1. The current responds with an exponential time constant T_1. Explicitly integrating equation (10), we find that the voltage which passes across the tip is given by

$$v_{s1}(t) = Z_1 I_t(t) = Z_1 \int_{-T_0}^{t} F_1(t') \frac{e^{(t'-t)/T_1}}{T_1} dt'. \tag{11}$$

This voltage becomes the incident pulse at the switch junction.

The Auston switch junction in Figure 4.2d is identical to the circuit in Figure 4.3b. The current through the switch therefore is determined by the equation

$$I + T(t-\tau)\cdot I' = F(t,\tau), \tag{12}$$

where

$$T(t-\tau) = \frac{R_s(t-\tau)\cdot(Z_1+Z_2)}{R_s(t-\tau)+(Z_1+Z_2)} C_s \text{ and } F(t,\tau) = 2\frac{v_{s1}(t) + T_s(t-\tau)\cdot v_{s1}'(t)}{R_s(t-\tau)+(Z_1+Z_2)}.$$

This equation is more complicated than the corresponding expression (equation (10)) for the tunneling current because the coefficient multiplying I' in the equation now varies with time.

The measured signal is given by the time-averaged value of the current I. One can write an analytical expression for the time-averaged current, but given a general shape for the response of the Auston switch, the expression for the time-averaged current is complicated and obtuse. The usual method of approximate solution follows the approach outlined in the discussion of Auston switches and involves expanding complicated exponential terms which arise. As discussed above, the expansion is not valid because of the large "on" conductance of the switch.

Instead, to gain some intuition into the nature of the measured signal, we will again consider a stepwise continuous R_s. In this case, R_s turns on at $t = \tau$, remains on for a time Δ, and then returns to its initial "off" state. With the assumption of a stepwise continuous R_s, equation (12) can be solved for the case of constant circuit components to deduce the current in each of the three regions $[-T_0,\tau]$, $[\tau,\tau+\Delta]$ and $[\tau+\Delta,\ T_0]$ [8]. Here $2T_0$ is the repetition period of the laser, and the regions are separated by the action of the switch.

The parameters T, T_s and R_s are constant in each region. The resistance of the switch drops from an "off" value of $R_{s\text{-}off}$ to an "on" value of $R_{s\text{-}on}$ during the interval $[\tau, \tau+\Delta]$. Correspondingly, T changes from T_{off} to T_{on}, T_s changes from $T_{s\text{-}off}$ to $T_{s\text{-}on}$, and F changes from F_{off} to F_{on}. To simplify the notation, I have dropped the "off" indices.

To write the solution at any point in time, we only need to keep track of the values of the circuit parameters and the initial conditions at the start of each of the intervals defined by the switch response. The solution for the current in each of the three regions is

$$I = \begin{cases} I_0 e^{-(t+T_0)/T} + \displaystyle\int_{-T_0}^{t} F(t') \frac{e^{(t'-t)/T}}{T} dt', & -T_0 < t < \tau \\ \left(I_0 e^{-(\tau+T_0)/T} + \displaystyle\int_{-T_0}^{\tau} F(t') \frac{e^{(t'-\tau)/T}}{T} dt' \right) \cdot e^{-(t-\tau)/T_{on}} + \displaystyle\int_{\tau}^{t} F_{on}(t') \frac{e^{(t'-t)/T_{on}}}{T_{on}} dt', & \tau < t < \tau + \Delta \\ \left[\left(I_0 e^{-(\tau+T_0)/T} + \displaystyle\int_{-T_0}^{\tau} F(t') \frac{e^{(t'-\tau)/T}}{T} dt' \right) \cdot e^{-\Delta/T_{on}} + \displaystyle\int_{\tau}^{\tau+\Delta} F_{on}(t') \frac{e^{(t'-(\tau+\Delta))/T_{on}}}{T_{on}} dt' \right] e^{-(t-(\tau+\Delta))/T} & \\ \qquad + \displaystyle\int_{\tau+\Delta}^{t} F(t') \frac{e^{(t'-(\tau+\Delta))/T}}{T} dt', & \tau + \Delta < t < T_0 \end{cases} \tag{13}$$

where $I_0 = I(-T_0)$.

Since the circuit is driven at the laser repetition rate, the system must reach a steady state in which $I_0 = I(T_0) = I(-T_0)$. We can use the solution of I for t in the region $[\tau+\Delta, T_0]$ to find $I(-T_0)$ and set this equal to I_0 to find the steady state solution

$$I_0\left(1-e^{(-2T_0+\Delta)/T}e^{-\Delta/T_{on}}\right)=$$

$$\left(\int_{-T_0}^{\tau}F(t')\frac{e^{(t'-\tau)/T}}{T}dt'\cdot e^{-\Delta/T_{on}}+\int_{\tau}^{\tau+\Delta}F_{on}(t')\frac{e^{(t'-(\tau+\Delta))/T_{on}}}{T_{on}}dt'\right)\cdot e^{-\frac{(T_0-(\tau+\Delta))}{T}}+\int_{\tau+\Delta}^{T_0}F(t')\frac{e^{(t'-T_0)/T}}{T}dt' \quad (14)$$

We measure the average current through the switch as a function of τ, given by

$$S(\tau)=\frac{1}{2T_0}\int_{-T_0}^{T_0}I(t,\tau)dt, \text{ or}$$

$$S(\tau)\approx\frac{1}{2T_0}\left[\int_{-T_0}^{\tau}I(t)dt+\int_{\tau}^{\tau+\Delta}I(t)dt+\int_{\tau+\Delta}^{T_0}I(t)dt\right] \quad (15)$$

This equation is valid when we observe the system over many laser repetition periods after the system has reached the steady state in which its initial conditions are identical for each repetition period. For the present circuit, the time constant of the system, on the order of Z_1C_s, is much shorter than the repetition period of the laser. The short time constant ensures that the system begins in the steady state in which $I_0 = 0$. However, the general formulation in which $I_0 \neq 0$ will be useful when we consider lumped circuits in the next section.

Plugging in the explicit form of $I(t)$ from equation (13) into each integrand and simplifying the resulting double integrals through integration by parts, we find

$$S(\tau)=\frac{1}{2T_0}\left[\int_{-T_0}^{T_0}F(t')dt'-\left(TI_0\cdot e^{-\frac{(T_0-\tau)}{T}}-\int_{-T_0}^{\tau}F(t')e^{\frac{t'-\tau}{T}}dt'\right)\left(1-\frac{T_{on}}{T}\right)\left(1-e^{-\frac{\Delta}{T_{on}}}\right)\right.$$

$$\left.+\int_{\tau}^{\tau+\Delta}\left\{F_{on}(t')\left(1-e^{\frac{t'-(\tau+\Delta)}{T_{on}}}\left(1-\frac{T}{T_{on}}\right)\right)-F(t')\right\}dt'\right] \quad (16)$$

We can now substitute the expression for I_0 in our expression for $S(\tau)$ to obtain, after some algebra,

$$S(\tau)=\frac{1}{2T_0}\left[\int_{-T_0}^{T_0}F(t')dt'+\int_{\tau}^{\tau+\Delta}\left(F_{on}(t')-F(t')\right)dt'\right.$$

$$-\left\{\int_{\tau}^{T_0}F(t')e^{\frac{t'-T_0}{T}}dt'\cdot e^{-\frac{T_0+\tau}{T}}+\int_{-T_0}^{\tau}F(t')e^{\frac{t'-\tau}{T}}dt'\right\}\frac{\left(1-\frac{T_{on}}{T}\right)\left(1-e^{-\Delta/T_{on}}\right)}{1-e^{-\frac{2T_0-\Delta}{T}}e^{-\Delta/T_{on}}} \quad (17)$$

$$\left.-\int_{\tau}^{\tau+\Delta}F_{on}(t')e^{\frac{t'-(\tau+\Delta)}{T_{on}}}\frac{\left(1-\frac{T}{T_{on}}\right)\left(1-e^{-\frac{2T_0-\Delta}{T}}\right)}{1-e^{-\frac{2T_0-\Delta}{T}}e^{-\Delta/T_{on}}}dt'+\int_{\tau}^{\tau+\Delta}F(t')e^{\frac{t'-(\tau+\Delta)}{T}}\frac{\left(1-\frac{T_{on}}{T}\right)\left(1-e^{-\Delta/T_{on}}\right)}{1-e^{-\frac{2T_0-\Delta}{T}}e^{-\Delta/T_{on}}}e^{-\frac{2T_0-\Delta}{T}}dt'\right]$$

This expression describes the time-averaged current of any circuit whose instantaneous current satisfies (12) for **generic** stepwise continuous T and F. The first two terms describe the signal which would be measured if the circuit had an infinitely fast response. We encountered these same terms in the capacitance-free circuit models.

The remaining terms in equation (17) are corrections to the signal due to the fact that the circuit does not respond instantaneously to changes in the driving term or switch resistance. The integral from τ to $\tau+\Delta$ involving F_{on} arises from current driven through the circuit by F_{on} while the switch is on and from charge draining off the

capacitance in the circuit after the switch is off. The integral from τ to $\tau+\Delta$ involving F results from discharge of the capacitance in the circuit while the switch is on. The first and second terms in the second line of equation (17) arise from buildup of steady-state initial current and current driven through the circuit while the switch is off, respectively.

The integrals in equation (17) in principle can be reduced to a simpler form when Δ and T_{on} are small, and T is either small or large when compared to the time scale for significant variations in the signal to be measured. In practice, however, the signal calculated through equation (17) is a tiny remainder which results from the subtraction of large, nearly equal terms. Therefore, one must be very, very careful in trying to reduce equation (17) to a simple intuitive result. It is much safer to solve the equation numerically.

There is, however, a general relation between the crashed and tunneling signals for the circuit we have considered in Figure 4.2d provided that the "off" decay time T is small compared to T_0 and that $R_{t\text{-}crash}C_t$ is small compared to risetimes in the signal being measured. Recalling the definitions of F, F_1, T and T_1 from equations (10) and (12), one can demonstrate that each term in equation (17) either involves an integration like

$$\int_{-T_0}^{\tau}\left\{\int_{-T_0}^{t}\left(\frac{v_i(t)+R_tC_tv_i'(t)}{R_t+\frac{1}{2}Z_0+Z_1}\right)\frac{e^{(t'-t)/T_1}}{T_1}dt'\right\}e^{\frac{(t-\tau)}{T}}dt$$

or can be reduced to an integration like the above. One then can tediously demonstrate through multiple integrations by parts that the terms involving v_i are related to the terms involving v_i' through the relation

$$R_tC_t\frac{d}{d\tau}A(v_i(t),\tau)=A(R_tC_tv_i'(t),\tau), \tag{18}$$

where A is an operator representing any of the various integrations which appear in (17). For example, in the simplest integration, the second term in (17), we find that

$$R_tC_t\frac{d}{d\tau}\int_{\tau}^{\tau+\Delta}\left\{\int_{-T_0}^{t}v_i(t)e^{\frac{(t'-t)}{T_1}}dt'\right\}dt=R_tC_t\left(\int_{-T_0}^{\tau+\Delta}v_i(t)e^{\frac{(t-(\tau+\Delta))}{T_1}}dt-\int_{-T_0}^{\tau}v_i(t)e^{\frac{(t-\tau)}{T_1}}dt\right)$$

$$=R_tC_t\left(\int_{\tau}^{\tau+\Delta}v_i(t)dt+\left\{\int_{-T_0}^{t}v_i(t)e^{\frac{(t'-t)}{T_1}}\Big|_{\tau}^{\tau+\Delta}-\int_{\tau}^{\tau+\Delta}v_i(t)dt+\right\}\right)=\int_{\tau}^{\tau+\Delta}\left\{\int_{-T_0}^{t}R_tC_tv_i'(t)e^{\frac{(t'-t)}{T_1}}dt'\right\}dt$$

satisfying equation (18). The last equality made use of the approximation $e^{-T_0/T}\approx 0$.

If $R_{t\text{-}crash}C_t$ is much smaller than the time scale of variations in $v_i(t)$, the dominant contribution to the crashed signal comes from terms involving integrals of $v_i(t)$. In tunneling range, the integrations of $R_tC_tv_i'(t)$ are no longer negligible. Equation (18) relates these terms to the terms measured from the crashed signal, and we recover the relationship postulated in Chapter 3

$$S_{tun}(\tau)\propto S_{cr}(\tau)+\frac{d}{d\tau}S_{cr}(\tau). \tag{19}$$

The proportionality constant is given by

$$\frac{R_t+\frac{1}{2}Z_0+Z_1}{R_{t-crashed}+\frac{1}{2}Z_0+Z_1}\approx\frac{R_t}{\frac{1}{2}Z_0+Z_1}. \tag{20}$$

This equation holds regardless of the values of R_s, C_s, R_{on}, Z_0, Z_1, Z_2, R_t and C_t provided $T << T_0$ and $R_{t\text{-}crash}C_t$ is much smaller than the time scale for variations in the signal. Since the Z_i's are likely much smaller than the R's,

$$T\approx(Z_1+Z_2)(C_s+C_t). \tag{21}$$

Even for values of Z_1+Z_2 on the order of $10^5\ \Omega$ and the sum of capacitances on the order of 1 fF, T is much smaller than T_0 (6.5 ns). Similarly, with $R_{t\text{-}crash}$ on the order of 100 Ω and C_t less than 1 fF, $R_{t\text{-}crash}C_t \le 0.1$ ps. Both conditions for the validity of equation (19) are satisfied.

In fact, even the assumption of a square pulse for the switch response is not necessary for equation (19) to be valid, provided the switch relaxes to its off state well before $t = T_0$. Additional steps in the response of the switch lead to additional double integrals which can be reduced to combinations of integrals in forms which satisfy (18). With an infinite number of steps, we can recover the exact solution for arbitrary switch response. Since each term satisfies (18), the relation between crashed and tunneling signals of equation (19) is valid for an arbitrary switch response.

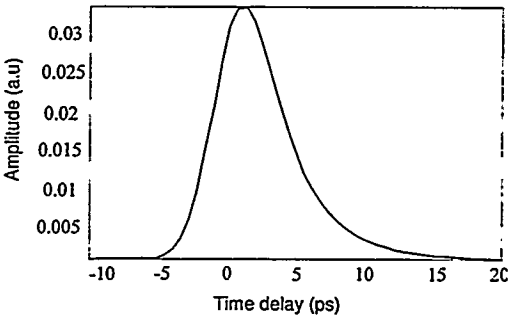


Figure 4.4. Numerical simulation of the crashed signal using equation (17) with F and T chosen to describe the circuit in Figure 4.2d. The tunneling signal exactly proportional to the crashed signal plus R_tC_t times its derivative.

A numerical simulation of the crashed signal for a 4 ps wide Gaussian voltage pulse on the sample is shown in Figure 4.4. The signal is qualitatively similar to those illustrated in Chapter 2. A width of 4 ps rather than 1 ps was chosen to account for dispersion on the tip.

The circuit model of Figure 4.2d that we have been considering describes some but not all of the features we observe in the tunneling signals. For example, the relationship described in equation (19) between crashed and tunneling signals does hold experimentally for long tips if we associate R_tC_t with the fitting parameter. If R_tC_t remains constant as R_t varies, the circuit predicts no change in the shape of the signal and a linear decrease in the amplitude with R_t as the separation between tip and sample is increased.. These are exactly the features we observe in the measured signal. However, the predicted amplitude of the signal is low by a factor on the order of 100, just as it was in the circuits of Figure 4.2a and b. Also, to obtain the correct amplitude ratio between crashed and tunneling signals, we require Z_1+Z_2 to have an unrealistically high value of ratio of 10^6 Ω.

As we found in Chapter 2, R_tC_t must be on the order of 4 ps to recover the shape of the signals we observe. This suggests an extremely small capacitance on the order of 0.1 aF which varies linearly with tunneling resistance and exponentially with tip-sample separation. This conclusion is fairly insensitive to changes in the values of the other circuit parameters (except R_t, of course).

4.2.3.2. Lumped circuit - Figure 4.2c

The effect of the transmission line in the previous model is to separate the switch and tunnel junctions so that the act of switching does not perturb the electromagnetic environment at the junction. When the STM junction and the photoconductive switch on the tip are close together, they can be considered to be two discrete elements in a lumped circuit. Changes in the switch resistance then do perturb the tunnel junction, altering the fraction of bias voltage dropped across the switch and altering the *RC* decay time of the whole circuit. During the time in which the switch is off, charge builds up on the capacitors in the circuit. When a probe laser turns the switch on, the capacitors rapidly discharge through the switch, altering the state of the circuit.

As in the previous example, I will consider a stepwise continuous switch response to facilitate a solution. For a sufficiently short "on" *RC* time this approximation should be fairly good. Applying Kirchoff's laws to the circuit in Figure 4.2c with constant circuit components, one finds that the current through the switch satisfies the now familiar equation

$$I + TI' = F(t,\tau) \qquad (22)$$

with

$$T = \frac{T_t \cdot R_s + T_s \cdot R_t}{R_s + R_t \cdot} = \frac{R_s R_t}{R_s + R_t \cdot}(C_s + C_t)$$

and

$$F = \frac{v_i(t) + (T_t + T_s)\cdot v_i'(t) \cdot + (T_t \cdot T_s)\cdot v_i''(t)}{R_s + R_t},$$

where $T_t = R_tC_t$ and $T_s = R_sC_s$.

These equations can be solved in each of the three regions delineated by the action of the switch in the same manner as in the previous section. Integrating the current over the interval $[-T_0, T_0]$ leads us again to equation (17). In the present case T is not necessarily small compared to T_0, and the self-consistency criterion imposed on the current I_0 is necessary. For a typical switch resistance of 100 MΩ and a typical total capacitance of 1 fF, T can be as

large as 100 ns, much larger than T_0, approximately 6.5 ns. In this case, one must be careful to measure the signal at each time delay continuously for on the order of 1 μs to avoid transient effects as the system reaches a steady state.

Equation (17) can be reduced to a simpler form with the approximations of small T_{on} and Δ and large T. As discussed in the previous section, these approximations can be quite misleading because the resulting signal arises from subtractions of large nearly equal terms. The equation again is most safely examined numerically.

Numerical simulations of the lumped circuit using equation (17) yield tunneling signals similar in shape and amplitude to the measured signals. The solid, dashed and dotted curves in Figure 4.5 illustrate three time-resolved tunneling signals calculated for tunneling resistances of 10^9, 10^8 and 10^7 Ω and scaled to a height of unity to enable direct comparison of their shapes. The fourth curve in Figure 4.5 represented by a dashed-dotted line is the calculated crashed signal for a junction resistance of 100 Ω. The voltage transient on the sample was taken to be a 2 ps Gaussian pulse with a peak amplitude of 100 mV. The capacitance of the switch was taken to be 1 fF, and the "on" and "off" resistances of the switch were 100 Ω and 100 MΩ, respectively. The width of the switch response was 1 ps. With these parameters the amplitude of the 10^8 Ω signal is approximately 10 pA. The calculated ratio of the amplitudes of the crashed signal to the tunneling signal at 1 GΩ, on the order of 10^4, is still much larger than we actually observe.

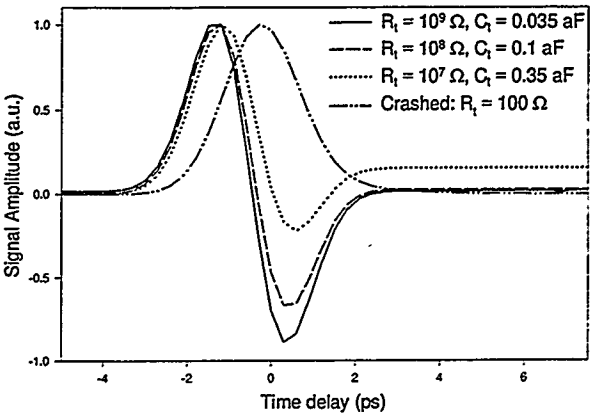


Figure 4.5. Signals predicted from a lumped circuit model including capacitances.

The amplitude of the calculated tunneling signal is sensitive to the capacitance of the tunnel junction. In order to obtain a signal strength on the order of 10 pA at a resistance of 10^8 Ω, the capacitance must be on the order of 0.1 aF. To reproduce the observed linear decrease in signal strength with R_t, C_t must vary roughly like $R_t^{-1/2}$. This variation in capacitance is exponential with distance, but the attenuation length is slower (roughly by a factor of 2) than that of the tunneling resistance.

The shapes of the tunneling signals are qualitatively similar to our measured signals. They precede the onset of the crashed signal and have narrower and in this case bipolar shapes. The shape now changes as we vary R_t, the negative dip becoming smaller as the resistance is decreased, in contrast with experimental observations. However, in our experimental measurements, we have not been able to vary the tunneling resistance over two the full orders of magnitude covered in the simulation.

4.2.3.3. Conclusions from circuit models

The results of the circuit models are at once exciting and discouraging. The fact that no simple circuit appears to completely model our results either for long tips or for short tips suggests that the measuring process is more complicated than we initially supposed. But the various models do agree on one key point: there must be a very small capacitance associated with the tunnel junction which varies inversely with the tunneling resistance.

The model of Figure 4.2d for long STM tips predicts the relationship we observe between the lineshapes of the crashed and tunneling signals. The model also predicts a linear decrease in the amplitude of the measured signal with tunneling resistance. In the context of this model, these observations require a tunnel junction capacitance on the order of 0.1 aF which varies inversely with the tunneling resistance. This suggests an exponential distance dependence which can not be described by a standard geometrical capacitance. The model fails to predict the magnitude of the observed amplitude for tunneling signals, and it requires a fairly large transmission line impedance to produce the correct ratio between the amplitudes of crashed and tunneling signals.

The lumped circuit model, which should apply well to short tips, predicts the correct amplitude for the signal we observe. In order to recover this amplitude, the capacitance again must be on the order of 0.1 aF. In order to recover the linear dependence of signal amplitude on tunneling resistance, the capacitance must vary approximately like the square root of the tunneling resistance. This variation, although slower than in the transmission line model, is still exponential with distance. The lumped circuit model also predicts a slight change in the shape of the tunneling signal as we change the separation between tip and sample, in contrast to observations.

4.3. Capacitance of the tunneling junction

The concept of macroscopic, geometrical capacitance assumes electrodes formed from perfect conductors which completely screen the electric field at their surfaces. For quasi-1-dimensional electrodes and at very small electrode separations, the screening length can become non-negligible, and the potential may not be uniform across the electrode surfaces. The presence of a highly localized tunneling current can further distort the potential distribution on the electrodes. These effects make the definition of capacitance difficult and should lead to discrepancies between the expected geometrical capacitance and the measured capacitance.

A second contribution to the capacitance across the junction comes from the action of tunneling itself. When the DC I-V curve across the junction contains a nonlinear region at the DC bias point, we expect a frequency dependent response to modulations at frequency ω once $\hbar\omega$ is larger than the width of the nonlinear region. A causal, finite, frequency dependent response in the resistance of the junction leads naturally to a reactance associated with the junction. There will also be reactances associated with the finite tunneling time across the junction and associated with the dynamics of formation of an image of the tunneling electron. Since the tunneling time and the plasma frequency in the electrodes are both on the order of femtoseconds, one expects these two processes to have a negligible effect on measurements of picosecond transient signals.

4.3.1. Geometrical capacitance

The geometrical capacitance of the junction can be estimated by considering the solution for a sphere (representing the tip) above an infinite plane (sample). This model is likely much more accurate than the simpler model of a plane-parallel capacitor. The capacitance of a sphere of radius R at a height z above a plane can be solved exactly with the method of images, yielding[9]

$$C_g = 4\pi\varepsilon_0 R \sum_{n=2}^{\infty} \frac{\sinh(\alpha)}{\sinh(n\alpha)}, \tag{23}$$

with

$$\alpha = \ln\left[1+\frac{z}{R}+\sqrt{\frac{z^2}{R^2}+2\frac{z}{R}}\right].$$

When R >> d, this expression can be simplified to[10]

$$C_g \propto \frac{1}{2}\ln\left(\frac{R}{2z}\right) \tag{24}$$

Thus the capacitance varies logarithmically with separation for the small separations and relatively large tip radii appropriate to STM. Figure 3.19 of Chapter 3 plots the capacitances calculated with equation (23) for tips with a 100 Å and a 1000 Å radius of curvature, respectively. These radii represent tips much sharper than those we routinely generate. It is clear that even with such sharp tips, the geometrical capacitance is larger and varies more slowly with distance than the capacitance deduced from analysis of our USTM measurements.

4.3.2. Effects from non-uniform potential distributions

4.3.2.1. Electrochemical capacitance

When the charge screening length in the electrodes becomes comparable to the spacing between the electrodes or comparable to the size of the electrodes, the capacitance must be defined as an electrochemical capacitance $C_\mu = e\frac{dQ}{d\mu}$, where Q is the charge on the capacitor plates and μ is the difference in the chemical potentials of the two plates. Buttiker has shown that for the simple case of a parallel plate capacitor, the capacitance is modified from $C=\frac{\varepsilon A}{d}$ to

$$C_\mu = \frac{\varepsilon A}{d+d_1+d_2}, \tag{25}$$

where d_1 and d_2 are on the order of the Thomas-Fermi screening lengths in the two conductors[11].

In an STM junction with a metallic tip and sample, the separation d and the screening lengths d_1 and d_2 all are on the order of a few angstroms. The effect of screening therefore should be to lower the measured capacitance by a

factor less than 10. The variation of the capacitance with distance also should be slower than that expected for geometrical capacitance. This is in contrast to our observations.

4.3.2.2. Distortions from localized tunneling current

Even if the electrodes were perfect conductors, the presence of a localized tunnel current between the two electrodes would change the potential distribution on the electrodes so that each no longer had a uniform potential across its surface. This is illustrated in Figure 4.6 for a stationary tunneling current. A potential gradient must exist along the surface to confine the current to the narrow region in which it tunnels.

Strong potential wells with widths on the order of the width of the tunneling current must exist at the spots where the tunneling current begins and ends[12]. These wells act to confine charge on the electrodes and therefore restrict capacitive effects to a very small electrode area while in tunneling range. A typical width for the tunneling current is on the order of 5 Å. Therefore, a crude estimate for the size of the tunneling capacitance based on a 5 x 5 Å plane parallel capacitor with 5 Å electrode separation is 10^{-20} F, consistent with the small capacitances we observe.

The strength of the gradient at the well boundaries will be proportional to the magnitude of the tunneling current. Hence the magnitude of confined charge and therefore the capacitance may be strongly dependent on changes in the tunneling current. However, it is difficult to deduce the exact dependence. Since the potential distribution is not uniform across the electrode surface, the definition of capacitance is again problematic, and one probably must use Buttiker's method from the previous section to make further headway..

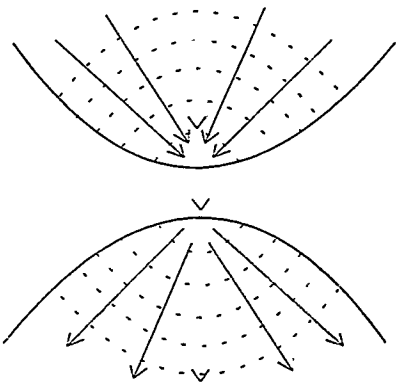


Figure 4.6. A graphical representation of the potential distribution which must exist to give rise to a local tunneling current. The dotted lines represent equipotential surfaces. The solid arrows represent current flow. The current flows along the gradients of the potential. At the surface of the electrodes, there must be a potential well (top electrode) or peak (lower electrode) with width on the order of the diameter of the tunneling current.

4.3.3. Photon assisted tunneling - Tucker model[13]

The Hamiltonian for a tunnel junction can be written in the form originally suggested by Cohen, Falicov, and Phillips[14]

$$H = H_L + H_R + H_T, \tag{26}$$

where

$$H_T = \sum_{lr\sigma} T_{lr}\, c_l^{\dagger} c_r + T_{lr}^{*}\, c_r^{\dagger} c_l$$

represents the tunneling interaction between left (L) and right (R) electrodes, and T_{lr} is the matrix element between states l and r on the left and right sides of the junction. The $c_{l(r)}^{\dagger}$'s and $c_{l(r)}$'s are the creation and destruction operators for particles on the left (right) sides of the junction.

The tunneling interaction is modified when one applies a time-varying potential V(t) to one side of the junction. The primary effect is to modify the phase factors of wavefunctions on that side. Assuming that the modulation is sufficiently slow that the system remains in thermal equilibrium and assuming that the modulation does not affect the electron wavefunctions in the tunnel barrier, linear response theory treating H_T as a perturbation leads to the expression

$$\langle I(t)\rangle = \mathrm{Im}\iint d\omega' d\omega W(\omega') W^{*}(\omega)\, e^{-i(\omega'-\omega)t}\, j\left(V_0 + \frac{\hbar\omega'}{e}\right). \tag{27}$$

In this expression, originally due to Werthamer[15], $j\left(\frac{\hbar\omega}{e}\right)$ is a response function which describes the tunneling current. W(ω) is the Fourier transform of the phase factor

$$\int d\omega' W(\omega') e^{-i\omega' t} = e^{\frac{-ie}{\hbar}\int dt'[V(t)-V_0]} \tag{28}$$

In the case of a time-independent potential, $W(\omega) = \delta(\omega)$, and therefore

$$I_{DC}(V) = Im\, j(V). \tag{29}$$

The imaginary part of the response function j thus is exactly the DC I-V curve. Hence Re j gives the reactive part of the tunneling current and must be related to Im j through a Kramers-Kronig relation.

Tucker evaluates equation (27) for a sinusoidal potential and DC bias $V(t) = V_0 + V_1 cos(\omega t)$ and finds that the

tunneling current can be described by

$$\langle I(t)\rangle = a_0 + \sum_{m=1}^{\infty}\left[2a_m \cos(m\omega t) + 2b_m \sin(m\omega t)\right], \qquad (30)$$

where

$$2a_m = \sum_{n=-\infty}^{\infty} J_n\left(\frac{eV_1}{\hbar\omega}\right)\left[J_{n+m}\left(\frac{eV_1}{\hbar\omega}\right) + J_{n-m}\left(\frac{eV_1}{\hbar\omega}\right)\right] I_{DC}\left(V_0 + \frac{n\hbar\omega}{e}\right) \qquad (31)$$

and

$$2b_m = \sum_{n=-\infty}^{\infty} J_n\left(\frac{eV_1}{\hbar\omega}\right)\left[J_{n+m}\left(\frac{eV_1}{\hbar\omega}\right) + J_{n-m}\left(\frac{eV_1}{\hbar\omega}\right)\right] \mathrm{Re}\,j\left(V_0 + \frac{n\hbar\omega}{e}\right). \qquad (32)$$

These equations show that the response can have resistive (cos) and reactive (sin) components at many harmonics of the driving frequency. When the photon energy is small compared to the width of nonlinearities in the DC I-V curve, equation (30) may be replaced by a classical modulation of the I-V curve. There is no classical analog for the reactive components; they vanish as the voltage width of I-V nonlinearities becomes large. The reactance arises from coupling between states of unequal energies on opposite sides of the junction. Tucker terms this coupling "quantum sloshing".

The reactive tunneling components are directly related to the Kramers-Kronig transform of the DC I-V curve. The I-V curve for STM can be approximated for voltages less than the work functions of the tip and sample by

$$I \approx A\cdot\left[3.16\cdot 10^{13}\frac{\sqrt{B}}{d}\cdot\left(V + \frac{3.27d^2}{B}V^3\right)\right]e^{-10.24d}, \qquad (33)$$

where A is the cross-sectional area (in m^2) of the tunneling current, B is the average height of the tunnel barrier expressed in volts, and d is the separation between tip and sample in nm[16]. We can use this expression to estimate the reactance of an STM tunnel junction[17]. Since the voltage and distance dependence are approximately separable, the Kramers-Kronig transform and summation in equation (32) will not affect distance dependence of the reactance. The reactance therefore will vary approximately exponentially with distance, as we observe experimentally.

Our picosecond transient pulses contain frequencies up to on the order of 1 THz, corresponding to a photon energy of 4 meV. This energy is small compared to the scale of the relatively weak nonlinearity in equation (33). Therefore we can use an approximation to (32) that Tucker derives for the small photon energy limit:

$$2b_m \approx -P\int_{-\infty}^{\infty}\frac{d\varpi}{\varpi}\int_0^{\pi}\frac{1}{\pi}d(\omega t)\cos(m\omega t)\left[I_{DC}\left(V_0 + V_1\cos(\omega t) + \frac{\hbar\varpi}{e} + \frac{m\hbar\omega}{2e}\right) - I_{DC}\left(V_0 + V_1\cos(\omega t) + \frac{\hbar\varpi}{e} - \frac{m\hbar\omega}{2e}\right)\right] \qquad (34)$$

Contributions to the reactance come only from the nonlinear regions of the I-V curve. (This is true even in equation (32)) It can be shown that the I-V curve does eventually become purely resistive, but the voltage at which this transition occurs is difficult to calculate[18]. A conservatively large value for this voltage, represented below by V_b, is 10 Volts. We can substitute (32) in (34) to obtain

$$2b_1 \approx -A\cdot\left[3.16\cdot 10^{13}\frac{\sqrt{B}}{d}\cdot\left(\frac{3.27d^2}{B}\right)\right]V_b e^{-10.24d}\cdot\frac{\hbar}{e}\cdot V_1\,\omega \qquad (35)$$

$$2b_n = 0, \quad n \neq 1$$

The reactive terms therefore contribute only a linear capacitance given by $C_{tunnel} = \frac{2b_1}{V_1\omega}$.

Referring to equation (32) and recalling that a typical tunneling resistance is 100 MΩ, we find

$$-A\cdot 3.16\cdot 10^{13}\cdot\frac{\sqrt{B}}{d}e^{-10.24d} \approx \frac{1}{R} = 10^{-8}\,\mathrm{Mhos}$$

Hence, for a tip-sample separation of 0.5 nm and a barrier height of 5 eV (approximately correct for metals) we find that $C_{tunnel} \approx 10^{-24}$ F. This value is much smaller than the capacitance we observe, but it has the correct exponential dependence.

Brandes and coworkers[19] extended the Tucker formalism to the case of tunneling through a 1-D channel, which

may be applicable to the STM junction. They considered the real part of the classical conductance of a circuit consisting of a 1-D channel with impedance Z_V in series with a tunnel junction. The junction consists of a capacitance C and a conductance Γ_T. The conductance of a solitary 1-D wire is $Z_V = \frac{h}{e^2}$. The conductance of the system then becomes

$$\Gamma(\omega) = \Gamma_T + \frac{h}{e^2} C^2 \omega^2 + O\left(\omega^4\right). \tag{36}$$

Brandes et al calculate the quantum mechanical response of such a system, obtaining

$$\Gamma(\omega) = \frac{e^2}{h} T(E_F) + \frac{e^2 m^2}{2\pi\hbar^3 d^2} \frac{k_F^2}{K_F^8} \omega^2 + O\left(\omega^4\right), \tag{37}$$

Here $T(E_F)$ is probability of transmission across the junction, so that $\frac{e^2}{h} T(E_F) = \Gamma_T$ is the conductance of the tunnel junction. The junction width is given by d, k_F is the Fermi wavevector, and K_F is the wavevector of an electron with energy V-E_F. A comparison of equations (36) and (37) shows that the system behaves as if it had an effective capacitance of

$$C = \frac{e^2 m}{2\pi\hbar^2} \frac{k_F}{K_F^4} \frac{1}{d} \tag{38}$$

For a gold tip and sample, V ≈ 10 eV, and E ≈ 5 eV. Assuming a tip-sample separation of 5 Å, the effective capacitance is on the order of 10^{-21} F. This is larger than the capacitance predicted by the simple Tucker theory but is still slightly smaller than our measured capacitances. Also, equation (38) predicts a linear variation with tip-sample separation. However, Brandes' results are based on the assumption that $k_F << K_F$. This condition that the Fermi level be substantially smaller than the barrier height above it applies to semiconductors but is inaccurate in the case of metals. A solution for the general case has not yet been published.

4.3.4. Coulomb blockade experiments

Some anomalous behavior in STM junction capacitance has previously been observed in STM junctions in DC experiments. These experiments study the phenomenon known as Coulomb blockade. A Coulomb blockade occurs in capacitive junctions small enough that the charging energy $E_C = e^2/2C$ associated with the addition of an electron to the capacitor plates exceeds the thermal energy kT[20]. The resistance across the junction must be larger than the quantum resistance h/e^2 in order that an electron's wavefunction can be considered to be localized on one side of the junction or the other. Additionally, the junction must be biased by a high impedance source (current biased); otherwise, quantum mechanical fluctuations wash out the charging effects[21].

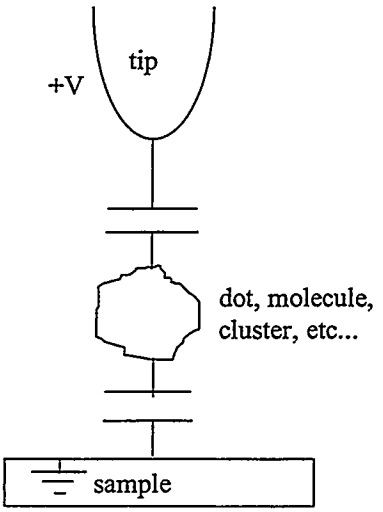


Figure 4.7. "Double junction" Coulomb blockade system. A small cluster between an STM tip and sample has capacitive paths

In practice it is extremely difficult to current bias a single junction. Therefore, to observe Coulomb blockade effects one usually studies a voltage biased double junction like that illustrated in Figure 4.7. The two capacitive junctions isolate a small central "dot" between two larger electrodes. If the resistances of the junctions are large and the capacitances are small, one observes a series of steps in the I-V curve each corresponding to the addition of a single electron to the dot.

STM experiments with a single **voltage** biased junction by van Bentum et al appeared to confirm the existence of a Coulomb blockade corresponding to a tip to sample capacitance on the order of 10^{-18}-10^{-19} F.[22] This magnitude agrees with our high frequency results. However, these results now are generally thought to have arisen from double junctions formed periodically as contaminant particles drift between the tip and sample[23].

A number of research groups have investigated capacitances in double junctions formed between the STM tip, small metallic clusters on an insulating layer on the sample, and the metallic sample substrate[24]. The capacitances inferred from these measurements are all on the order of 10^{-18} F or less and vary linearly with tip-sample separation. The small size of the capacitances is not so surprising, because the clusters which form the "dots" in the system are

typically on the order of 50 nanometers or less in diameter. However, the linear dependence on distance is surprising from the standpoint of geometrical capacitance. The decay should be closer to a logarithmic dependence as noted in the discussion above. There are as yet no published reports of direct measurements of the capacitance in an STM junction.

Chapter 4 References:

1. Daniel R. Grischkowsky, Mark B. Ketchen, C.-C. Chi, Irl N. Duling, III, Naomi J. Halas, Jean-Marc Halbout, and Paul G. May, "Capacitance Free Generation and Detection of Subpicosecond Electrical Pulses on Coplanar Transmission Lines", *IEEE J. Quant. Elec.*,**QE-24**. p. 221 (1988).
2. The series inductance of the junction is on the order of the internal inductance of a thin wire 10 Å long, or 10 10^{-10} $\mu_0/8\pi$ = 5 10^{-17} H. The impedance at THz frequencies is on the order of 100 $\mu\Omega$, utterly negligible when compared to the tunneling resistance or when compared to the lowest reasonable estimate for the impedance of the capacitive path. The inductance of the switch is also expected to be small. The only point in the circuit where a larger inductance might occur is at the sharp bend which marks the connection of the tip to the switch electrode. Even at this point, the inductance is expected to be negligible.
3. Even the calculation of dispersion on transmission lines with simple well-defined geometries is not trivial. See, for example G. Hasnain. A. Dienes, and J. R. Whinnery "Dispersion of Picosecond Pulses in Coplanar Transmission Lines", IEEE Trans. Microwave Theory Tech., **MTT-34**, p 738 (1986).
4. A review of the physics and properties of fast photoconductors can be found in D.H. Auston, "Picosecond Photoconductors: Physical Properties and Applications", in *Picosecond Optoelectronic Devices*, Academic press, San Diego, (1984).
5. The change in dielectric constant can be determined through the Kramers-Kronig relations. It is probably smaller than 10% even for carrier densities on the order of $10^{19}/cm^3$. See, for example D. Botkin, S. Weiss, G. Sucha, D.S. Chemla and J.M. Wiesenfeld, "Ultrafast dynamics of the optical mode of a 1.5 μm multiple quantum well optical amplifier", *Appl. Phys. Lett.*, 64, p. 2861 (1994), and the references therein. These papers discuss the changes induced at optical frequencies. At GHz-THz frequencies, the effect should be even smaller.
6. The discussion in this section mostly follows: David H. Auston, "Impulse Response of Photoconductors in Transmission Lines", *IEEE J. Quant. Elec.*, **QE-19**, p. 639 (1983). It also results in part from helpful discussions with Ajay Nahata.
7. The time delay associated with propagation on the transmission line has also been neglected because it merely adds an uninteresting uniform shift in time.
8. The starting point for this lumped circuit solution was suggested by Stephan Glutsch. He noted that the circuit could be solved for an arbitrary switch response by solving a static circuit over a brief interval, determining the current at the endpoint of the interval, and using this current as the initial condition for the next time interval. This solution converges to the exact solution as the width of the intervals goes to zero. The solution is also exact for a stepwise continuous switch response like the one I consider.
9. Dror Sarid, *Scanning Force Microscopy*, Ch. 11, Oxford University Press (1991).
10. Stephan Glutsch, *private communication*
11. M. Büttiker, "Capacitance, admittance, and rectification properties of small conductors", *J. Phys.: Condens. Matter*, **5**, p. 9361 (1993).
12. B. Laikhtman and E. L. Wolf, "Tunneling Time and Effective Capacitance for Single Electron Tunneling", *Phys. Lett. A*, **139**, p .257 (1989).
13. This section closely follows the treatment of J.R. Tucker, "Quantum Limited Detection in Tunnel Junction Mixers", *IEEE J. Quantum Electron.*, **QE-15**, p .1234 (1979) and J.R. Tucker and M.J. Feldman, "Quantum detection at mm wavelengths", *Rev. Mod. Phys.*, 57, p. 1055 (1985).
14. M. H. Cohen, L. M. Falicov and J. C. Phillips, "Superconductive tunneling", *Phys. Rev. Lett.*, 8, p316 (1962).
15. N. R. Werthamer, "Nonlinear Self-Coupling of Josephson Radiation in Superconducting Tunnel Junctions", *Phys. Rev.*, **147**, p 255 (1966).
16. E. Clayton Teague, "Room Temperature Gold-Vacuum-Gold Tunneling Experiments", *J. Res. Nat. Bur. Stand.*, **91**, p. 171 (1986).
17. We cannot accurately measure the I-V response of the STM since the photoconductive switch is in series with the tip.
18. John G. Simmons, "Generalized Formula for the Electric Tunnel Effect between Similar Electrodes Separated by a Thin Insulating Film", *J. Appl. Phys.*, **34**, p 1793 (1963).
19. T. Brandes, D. Weinmann and B. Kramer, "Frequency-dependent conductance of tunnel junctions", *Europhys. Lett.*, **22**, p. 51 (1993).
20. D. V. Averin and K. K. Likharev, "Coulomb Blockade of single-electron tunneling, and coherent oscillations in small tunnel junctions", *J. Low Temp. Phys.*, 62 p 345 (1986)
21. H. Grabert and M. H. Devoret (Eds.), *Single Charge Tunneling*, NATO ASI Series B, Vol 294, Plenum, New York (1992).
22. P. J. M van Bentum, H. van Kempen, L. E. C van de Leemput, P. A. A Teunissen, "Single-electron tunneling observed with point-contact tunnel junctions", *Phys. Rev. Lett.*, **60**, p 369, (1988).
23. Z. Y. Rong, L. F. Cohen, and E. L. Wolf, "Scanning Tunneling Microscope Observations of the Coulomb Blockade", *Phys. Lett. A*, **146**, p 281 (1990).
24. See, for example, D. Anselmetti, T. Richmond, A. Baratoff, G. Borer, M. Dreier, M. Bernasconi, and H-J Guntherodt, "Single Electron Tunneling at Room Temperature with Adjustable Double Barrier Junctions", *Europhys. Lett.*, **25**, p 297 (1994), and references therein.

Chapter 5
PROSPECTS FOR THE FUTURE

There are two main avenues to pursue for future work with the USTM: single point measurements and time resolved images, or movies. Measurements at a single point can be utilized to study the physics of high frequency transport in 1-d constrictions. Such measurements also will help to elucidate the nature of the response of the USTM as an instrument. An important ingredient in the interpretation of single point measurements is a reliable theoretical model for the USTM.

The prospect of time resolved movies of surface dynamics on an atomic scale is extremely exciting. It should be possible to spatially and temporally resolve phenomena such as vibronic motion and charge transfer. It may eventually be possible to observe electronic or conformational changes in chemical reactions occurring on surfaces. The interpretation of ultrafast movies does not necessarily require a complete physical understanding of single point measurements. Experiments that look for spatial features which change as a function of time should be relatively insensitive to the details of time-resolved tunneling. An example of such a feature is the path of the peak amplitude of a scattered surface distortion. Another example is the final orientation or position of a charge density wave relative to the underlying atomic lattice after excitation in the presence of a pinning impurity.

The purpose of this chapter is to briefly illustrate some possible future experiments and some experiments now in progress with the USTM.

5.1. Measurements at a single point

5.1.1. δz response

Our measurements have characterized the response of the USTM to an applied voltage impulse $\delta v(t)$. Since the STM also is sensitive to topographic signals, an obvious next step is to measure the response of the USTM to a topographic impulse $\delta z(t)$. This type of experiment serves as a prelude to time-resolved imaging of phonon propagation across interfaces or surface structures.

Since picosecond packets of phonons are strongly attenuated after propagation over several hundred Å, the impulse must be generated close to the tip. However, bringing the pump laser beam close to the tip leads to the heating problems discussed in Appendix 1. An elegant solution[1] is to generate a packet of phonons on the back side of a thin metallic film and to measure the packet after propagation to the STM tip on the front of the film. Our microscope is designed specifically to allow such backside excitation of the sample. Similar experiments have been performed with photothermal deflection detection of the phonon packet[2]. However, the measurements are complicated by transient changes in the reflectivity of the surface not related to surface motion.

A circuit model for the USTM experiment can be set up in the same manner as for the electronic measurement. In the δz measurement, the applied voltage is constant, while the resistance and capacitance of the junction change. Hence this measurement, in addition to providing information on phonon propagation through a thin film, can be used to study the capacitance of the junction.

5.1.2. Capacitance

We also can study the capacitance in more direct ways. One possibility is to explore the dynamic response of the Coulomb blockade system described in the previous chapter. For an ultrafast measurement, one simply prepares metal clusters over a thin oxide on one electrode of a transmission line. A voltage pulse on the transmission line produces an ultrafast I-V curve. The charging energy of the cluster should lead to steps in time in the time-resolved trace just as it leads to steps in voltage in the DC I-V measurement. From the step spacing, we can extract the capacitances from the tip to the cluster and from the cluster to the sample. There also may be other interesting dynamics in the system, such as vibrational modes in the cluster.

A single capacitive junction biased by a current source has been predicted to give rise to a phenomenon known as time-correlated tunneling[3]. Time-correlated tunneling never has been observed experimentally. This phenomenon arises from the fact that the charge on the electrodes can change continuously, while a tunneling event provokes a sudden discontinuous drop of magnitude e. The current source causes a linear increase in the charge across the tunnel junction until the charging energy makes a tunneling event energetically favorable. An electron tunnels across the junction, reducing the charge on the electrodes by e, and the process repeats itself. Time-correlated tunneling therefore leads to a sawtooth tunneling current with frequency I/e. A current of 10 nA leads to a 16 ps period, well within our resolution.

The currents we have measured with the USTM tend to be smaller, on the order of 1 nA, corresponding to a

period of 160 ps. Since we typically observe signals over a range of a few tens of picoseconds, it would be difficult to detect oscillations arising from time correlated tunneling in the measurements of Chapter 3. More importantly, we apply a voltage rather than current bias across the junction and therefore should not produce these oscillations. One possible way to "current bias" the junction for future experiments is to fabricate a highly resistive lead in series with the junction, either on the sample or the tip.

5.1.3. Point contact measurements

The DC transport in quantum point contacts in a two-dimensional electron gas is well understood. The DC transport in the atomic sized contact formed when the STM tip touches the sample surface is less well understood but has been the subject of much recent work[4]. High frequency transport in quantum point contacts in a two dimensional electron gas also has been studied recently[5]. High frequency transport in an atomic sized point contact is not well understood and has not been studied at all. At present the USTM is the only tool available to study such a system.

One interesting experiment is to study the change in the USTM response as the tip moves from tunneling range into ohmic contact with the sample. If the STM is stable enough, a point contact can be formed in the regime intermediate between crashed and tunneling. We have performed some preliminary measurements of this sort using the USTM in air. Ultimately, this type of experiment should be performed in vacuum, where tip and sample can be clean and well characterized.

5.1.4. Integrated circuits

An important technological application of USTM is integrated circuit testing. We have demonstrated on the order of $1\ \mathrm{mV}/\sqrt{\mathrm{Hz}}$ sensitivity in the measurement of repetitive picosecond signals. The USTM can locate and measure signals at very small test points. It also can measure signals as they propagate on the metal lines of the circuit. Some additional preparatory experiments must be done before the USTM is suitable for IC testing, however. One should measure voltage pulses much longer than 1 picosecond to ensure that we can easily interpret the resulting waveforms. One also should measure sets of closely spaced pulses to see how applicable the measurement is to the study of rapid digital signals.

5.2. Movies

5.2.1. Introduction

Ultrafast movies can be produced in at least two ways. The first is to measure single point time resolved traces at a series of desired spatial pixels. The second is to measure the spatially resolved image arising at fixed time delay between pump and probe beams for several different time delays. Each spatially resolved image then becomes one frame of the time-resolved movie.

Signal to noise levels currently make generation of a time resolved movie extremely difficult. Another important concern is how one can excite the sample locally (on a nanometer scale) with a laser beam which has on the order of micron spatial extent. One solution is to choose a sample which itself defines the spatial size of the excitation. For example, the vibrations of an adsorbed molecule on the surface are localized even though the effective area of the impulse which excites it may be large.

5.2.2. Noise

The biggest obstacle to time and space resolved measurements is our present signal to noise levels. To make a time resolved measurement with a signal strength equivalent to our current signal strength requires at least 30 seconds integration at each point in the image. Even with a small number of spatial pixels, say 16 x 16, the movie would take on the order of 100 minutes to acquire. In ambient conditions the STM is not stable enough to make such an image over an area of a few nanometers without substantial drift and low frequency noise over the period of the measurement.

Noise levels can be attacked on several fronts. Signal to noise ratios can be increased by reducing the bandwidth of the measurement through lock-in detection or by increasing the fast scanning frequency. The differential time measurement technique discussed in Chapter 2 is an example of this approach.

Increasing the stability of the microscope also can lead to dramatic improvements. Increased vibration isolation, operation in partial vacuum or ultrahigh vacuum, and possibly operation at low temperature should reduce the noise of the USTM by a large factor. One can also attempt to detect and correct for drift to allow much longer measurement times. Such a scheme can be accomplished through real time image processing.

5.2.3. Semiconductor devices

A semiconductor heterostructure is an example of a system in which the sample structure can determine spatial resolution. For example, a heterostructure consisting of several 10 nm wide quantum wells can be illuminated with a laser of energy high enough to excite carriers in the barrier regions. The carriers diffuse and fall into the wells on a time scale of several picoseconds. We should be able to observe this diffusion and carrier capture by monitoring changes in the contrast between images taken at different time delays. This diffusion and capture, which influences the performance of semiconductor lasers, has been indirectly measured in optical experiments in semiconductor optical amplifiers[6].

Another interesting system is a p-n junction illuminated by a short pulse laser. In this case, the excited carriers can be expected to screen the fields across the junction, resulting in an initial broadening of the depletion region. As electrons and holes diffuse in opposite directions at different speeds, the depletion region should move laterally and then return to its initial position as the system relaxes.

The laser can excite on the order of 10^{18} carriers/cm^3 in the system. An STM can detect such changes in carrier density at p-n junctions and also on semiconductors irradiated with light[7]. An advantage of systems like these is that spatial changes occur mainly in the dimension perpendicular to the plane of the interfaces. Hence, the USTM movie need only be made in one spatial dimension dramatically reducing the time needed for a measurement.

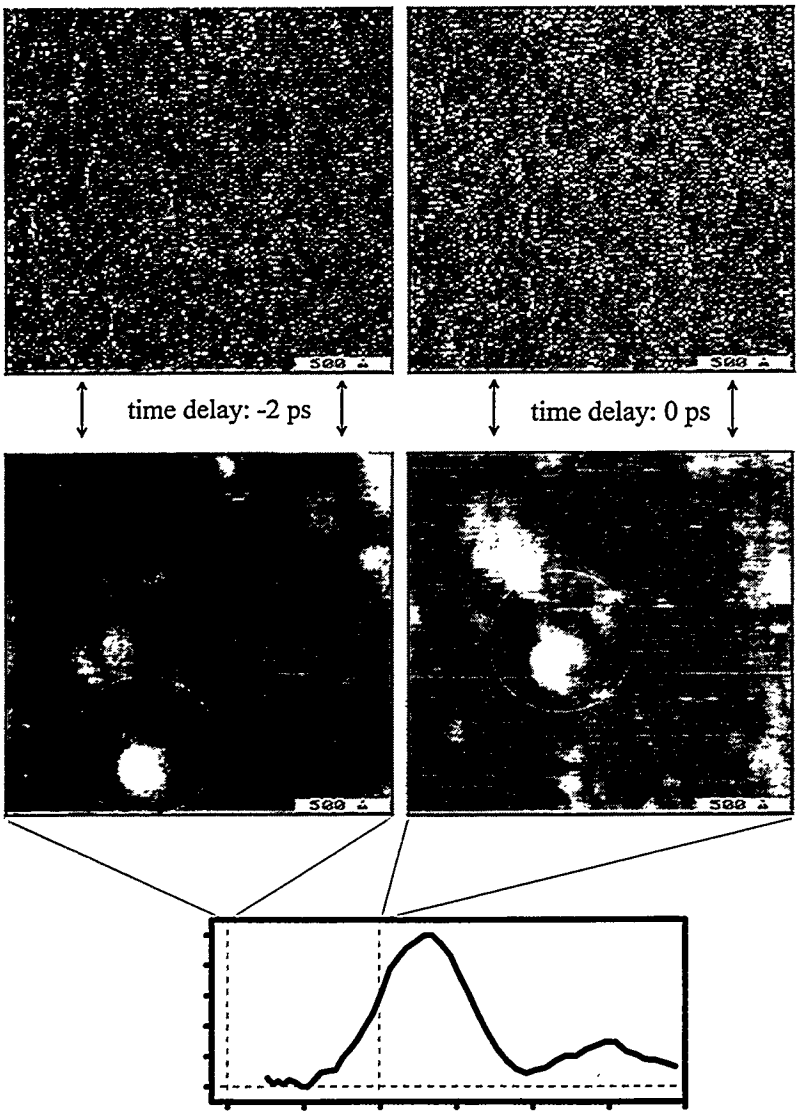


Figure 5.1. Two frames of a movie. 2000 x 2000 angstrom topographic and differential time delay images are shown at time delays of -2 ps and 0 picoseconds, respectively. The single point time resolved tunneling current appears below. Dashed vertical lines represent the temporal positions of the images.

5.2.4. A "B" movie ? The voltage pulse on a transmission line

The USTM can make images over windows on the order of 1 μm x 1 μm or smaller. The spatial extent of the voltage pulse measured in Chapter 3 is much larger, on the order of 100 μm wide. We therefore expect an ultrafast movie of the pulse to reveal only a uniform rise and fall in the height of a topographic image as time delay changes. This represents a movie with a rather dull story line.

In fact the story line may be slightly more engaging. One expects that the fields generated by the voltage pulse at sharp features on the surface will be different than the corresponding fields at smooth regions of the surface. Images generated from the slope of the I-V curve at each pixel reveal such an effect. The response of the USTM to a macroscopically large voltage pulse therefore also should exhibit spatial features which correspond to surface topography.

These features cannot be discerned in a standard time resolved measurement due to excessive noise. To reject all signals not associated with the pump laser, we use the differential time delay method discussed in Chapter

2, modulating the time delay of the pump pulse and using lock-in detection at the modulation frequency. Figure 5.1 shows topographic and differential time delay images measured at a time delay of -2 picoseconds and 0 ps. The topographic and differential signals were measured simultaneously. The corresponding single point time resolved tunneling current is shown below the images. Zero time delay corresponds to a point halfway up rise to the initial peak in the single point measurement. The differential signal should be strongest at this point; it should be zero at -2 ps.

The images are 128 x 128 pixels and cover 2000 x 2000 Å. The time delay was modulated at 360 Hz, and the signal was measured for 15.6 ms per pixel. The lock-in time constant was 1 ms. The tunneling resistance was 130 MΩ, and the DC current was 1 nA. The differential time delay images are rendered with a light source at the side of the image. This tends to enhance vertically aligned features in the images. The topographic signals are displayed with height coded in grayscale.

The differential time delay images are noisy, but one sees clearly features which correspond to topography on the sample. An illustrative feature is tracked through the images with a white circle. (The vertical displacement of the circled feature in the -2 ps and 0 ps images results from slow drift of the tip relative to the sample.) The features appear even at negative time delay, which suggests that they are at least partly due to artifacts in the measurement. However, the contrast in the 0 ps differential time delay image is noticeably enhanced relative to the -2 ps image, as we expect for time resolved features. The resolution in the differential time delay image is approximately 50 Å, limited by pixel size. These images may be the first direct evidence of nanometer spatial resolution in time resolved measurements with the USTM.

This is a very preliminary result. Much work still needs to be done to improve the quality of the images. Drift rates are too large to measure more than a few frames. Apparent artifacts at negative time delay need to be understood and removed, if possible. One way to do this is to measure the signal at each pixel with and without the pump laser on, and to subtract the two images. In practice, one would do the subtraction line by line to mitigate the effects of drift. This technique and the improvements discussed above are currently being added to the USTM.

5.3. Summary

My dissertation has discussed the concept, design, construction, and implementation of the ultrafast STM. This invention represents a dramatic leap over prior work in the field of time-resolved STM and has generated a surge of related research. Our instrument has demonstrated sub-picosecond temporal resolution, improving the time resolution attainable with STM by at least 6 orders of magnitude. Although we do not have a clear direct measurement of the spatial resolution of the USTM, indications are that it is better, and possibly much better, than 100 Å.

I have used the USTM to study the high frequency response of the STM tunnel junction. Physically reasonable circuit models for the system partially describe my experimental results. An analysis of the USTM measurements with these models suggests that the tunnel junction capacitance is several orders of magnitude smaller and varies much more rapidly than one expects from geometrical considerations. I have outlined some possible explanations for this behavior. However it is clear that more work needs to be done to fully understand the mechanisms involved in the measurements. I have sketched some of the many possible directions for future research.

The USTM is the first widely applicable tool to enable the observation of phenomena and processes on a picosecond and nanometer scale. However, noise performance still must be improved substantially before one can routinely generate ultrafast movies. Ultimately, this capability should open the door to diverse applications in the fields of physics, chemistry and materials science.

Chapter 5 References:

1. Shimon Weiss originated the concept of a Δz measurement to complement ΔV measurements and also suggested the specific experiment discussed here.
2. O.B. Wright and K. Kawashima, "Coherent phonon detection from ultrafast surface vibrations", *Phys. Rev. Lett.* **69**, p 1688 (1992).
3. K. K. Likharev, "Correlated discrete transfer of single electrons in ultrasmall tunnel junctions", *IBM J. Res. Devel.*, **32**, p 144 (1988).
4. N. Agrait, J.G. Rodrigo, and S. Vieira, "Conductance steps and quantization in atomic size contacts", Phys. Rev. B, **47**, p 12345 (1993). L. Olesen, E. Laegsgaard, I. Stensgaard, F. Besenbacher, J. Schiotz, P. Stoltze, K. W. Jacobsen, and J. K. Norskov, "Quantized conductance in an atom-sized point contact", *Phys. Rev. Lett.*, **72**, p 2251 (1994).
5. C. Karadi, S. Jauhar, L. P. Kouwenhoven, K. Wald, J. Orenstein, P. L. McEuen, Y. Nagamune, and H. Sakaki, "Dynamic response of a quantum point contact", *J. Opt. Soc. Amer. B*, **11**, p. 2566 (1994).
6. J. M. Wiesenfeld, S. Weiss, D. Botkin, D.S Chemla, "Carrier transport effects and dynamics in multiple quantum well optical amplifiers", *Opt. Quant. Elec.*, **26**, p S731 (1994)
7. R. M. Silver, J. A. Dagata and W Tseng, "Ambient and vacuum scanning tunneling spectroscopy of sulfur- and oxygen-terminated gallium arsenide", *J. Appl. Phys.*, **76**, p5122 (1994).

Appendix 1
ON ULTRAFAST MEASUREMENTS USING INTRINSIC STM NONLINEARITIES

The native nonlinearities available for exploitation in the STM junction are the *I-V* curve and the *I-Z* curve. In this section I analyze four hypothetical ultrafast experiments: the measurement of electronic or topographic signals via modulation of the *I-V* curve or via the *I-Z* curve. The discussion assumes that picosecond modulation along of one of these curves has somehow been accomplished with the probe laser pulse. Such modulation is, in fact, difficult or impossible to achieve in practice.

A1.1. Theory

A1.1.1. Circuit models

As discussed in Chapter 4, the STM junction is assumed to consist of a parallel combination of resistance and capacitance. We found in Chapter 1 that the *RC* time constant defined by the spreading resistance, on the order of 10 fs, is small compared to the temporal width of the incident signals from the pump and probe. A modulation or an electronic process which changes the effective bias across the STM junction will change the resistance of the tunnel junction in accordance with the nonlinear *I-V* curve. A modulation which changes the separation between tip and sample or a process which produces topographic changes in the sample will change the resistance of the junction in accordance with the nonlinear *I-Z* curve and also will change the capacitance of the junction.

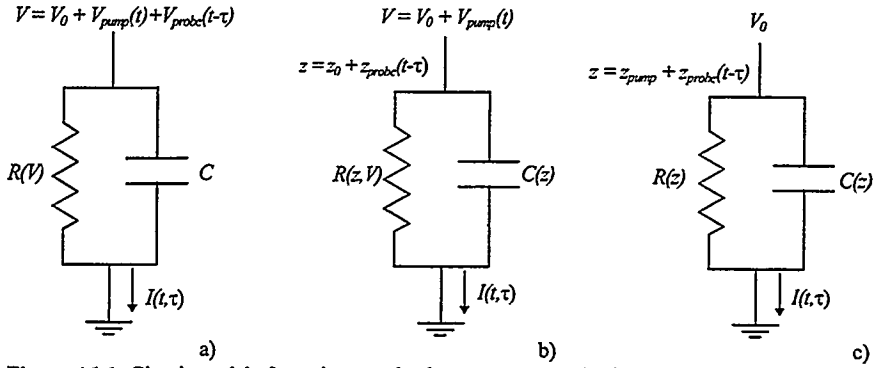


Figure A1.1. Circuit models for a time resolved measurement : a) of electronic variation using the STM's I-V relation. b) of electronic variation using the I-Z curve, or, with the roles of pump and probe reversed, of topographic variation using the I-V relation, and c) of topographic variation using the I-Z relation.

Figure A1.1 shows circuit models for each of the hypothetical measurements. Since all the "action" occurs at the STM junction, circuit models should describe the system quite well. Figure A1.1a illustrates the measurement, via probe induced *I-V* modulation, of a pump induced electronic process which changes the bias across the junction. Figure A1.1b illustrates a measurement of the same process via *I-Z* modulation. If the subscripts "pump" and "probe" are switched, Figure A1.1b illustrates the measurement of transient topographic changes via *I-V* modulation. Figure A1.1c illustrates the measurement of topographic changes via I-Z modulation.

A1.1.2. Calculated signals

The most straightforward example of a time-resolved measurement using the STM's *I-V* curve is a measurement of a pump-induced voltage, V_{pump}, on the sample, as in Figure A1.1a. The probe is assumed to generate an additional voltage, V_{probe}, across the tunnel junction. The temporal position of this probe pulse moves with the time delay, τ, between pump and probe beams. The junction is modeled as the parallel combination of a nonlinear resistor and a linear capacitor biased by the pump and probe generated voltages and by the STM's usual DC bias. Simple circuit analysis leads to the result

$$I(t,\tau) \approx \frac{V_0 + V_{pump}(t) + V_{probe}(t-\tau)}{R\left(V_0 + V_{pump}(t) + V_{probe}(t-\tau)\right)} + C \cdot \left\{V'_{pump}(t) + V'_{probe}(t-\tau)\right\},$$

where τ is the time delay between pump and probe. Only terms which multiply the two voltages can generate τ-dependent terms in the time-averaged signal. Hence the second term, which involves the junction capacitance, plays no role, and the signal is seen to arise solely from ohmic current through the tunneling resistance.

For small V_{pump} and V_{probe}, the measured signal can be expressed as

$$S(\tau) = \frac{1}{2}\frac{\partial^2 I}{\partial V^2}(V_0) \cdot \frac{\int_{-T_0}^{T_0} V_{pump}(t) \cdot V_{probe}(t-\tau)\,dt}{2T_0} + \text{constant terms}.$$

This expression is proportional to a convolution of the pump-induced voltage response with the probe induced response. The convolution measures the average of the pump voltage over a window centered at time τ and defined by the shape of the probe voltage. In an ideal case, the probe voltage is a δ-function with amplitude V_{probe}, and the measured signal becomes

$$S(\tau) = \left(\frac{1}{2} \frac{\partial^2 I}{\partial V^2}(V_0) \cdot \frac{V_{probe}}{2T_0} \right) \cdot V_{pump}(\tau) + \text{ constant terms}. \tag{1}$$

In this ideal case, the measured signal is exactly proportional to the height of the voltage pulse we wish to measure at time τ. When the probe pulse has a finite width, the signal measured as a function of time delay τ is blurred by the width of the probe pulse. The amplitude of the signal depends on the curvature of the *I-V* curve and the amplitude of the probe response and is reduced by the duty cycle of the measurement.

The analysis becomes slightly more complicated when the pump-induced signal changes the width of the tunnel gap. The circuit model for this case is illustrated in Figure A1.1b with the roles of pump and probe switched. The pump-induced process on the sample (for example a propagating packet of phonons) produces changes both in the resistance and in the capacitance of the junction. A simple analysis of the circuit leads to

$$I(t) = \frac{V_0 + V_{probe}(t-\tau)}{R\left(z_0 + z_{pump}(t), V_{probe}(t-\tau)\right)} + \frac{d}{dt}\left\{ C(z_0 + z_{pump}(t)) \cdot \left(V_0 + V_{probe}(t-\tau)\right) \right\}$$

The term involving the capacitance is the time derivative of *CV*. If the system relaxes to its initial state between sets of laser pulses, so that $z_{pump}(T_0) = z_{pump}(-T_0)$, then this term integrates to zero in the time-averaged signal. Again the measured signal arises from ohmic currents across the STM junction. The measured signal is

$$S(\tau) = \frac{1}{2T_0} \int_{-T_0}^{T_0} \left\{ \frac{V_0 + V_{probe}(t-\tau)}{R\left(z_0 + z_{pump}(t), V_{probe}(t-\tau)\right)} \right\} dt,$$

which reduces to

$$S(\tau) = \frac{\partial I}{\partial V}(V_0, z_0) \cdot \frac{\partial I}{\partial z}(V_0, z_0) \cdot \frac{\int_{-T_0}^{T_0} z_{pump}(t) \cdot V_{probe}(t-\tau) dt}{2T_0} + \text{ constant terms},$$

for $\frac{z_{pump}}{z_0}$ and $\frac{V_{probe}}{V_0}$ much smaller than unity.

As before, we measure a convolution of the probe voltage pulse with the amplitude of the process we wish to measure on the sample. The convolution represents an integral of the process we are interested in on the sample over a window around time τ with a width given by the temporal width of the probe voltage pulse. For sufficiently narrow probe pulses, the measured signal is proportional to $z_{pump}(\tau)$. The strength of the measured signal is scaled by the a second partial derivative of the *I(V,z)* response and by the amplitude of the probe voltage pulse and is reduced by the duty cycle of the measurement.

Similar analyses can be performed for measurements made when the probe modulates the STM *I-Z* curve. The signal arising from a measurement of voltage on the sample is the same as the previous expression with the roles of pump and probe reversed. The measurement of a topographic signal using the STM I-Z curve, as illustrated in 1c, reduces to

$$S(\tau) = \frac{\partial^2 I}{\partial z^2}(V_0, z_0) \cdot \frac{\int_{-T_0}^{T_0} z_{pump}(t) \cdot z_{probe}(t-\tau) dt}{2T_0} + \text{ constant terms},$$

for small z_{pump} and z_{probe}.

The time resolution of each of these measurements is determined by the temporal width of the probe-induced modulation. As discussed in Chapter 1, the time resolution is ultimately limited by the largest of the junction *RC*, the tunneling time, and the inverse of the plasma frequency to on the order of 10-100 fs. Because only signals arising from ohmic currents survive in the time averaged measurements, the spatial resolution of these measurements should

be equal to the spatial width of the tunneling current.

A1.2. Practice

The signals which arise from the hypothetical measurements above are exactly what one hopes for in a pump-probe experiment: a direct convolution of the pump-induced sample response with a probe-induced gating modulation. In practice, serious difficulty lies in how to modulate *I-V* or *I-Z* on a picosecond scale with the probe laser pulse. As discussed in the introduction to the thesis, Nunes and Freeman have successfully modulated the *I-Z* curve in the nanosecond regime. However, it is extremely difficult to modulate motion, even angstrom scale motion, in a well characterized way in the picosecond regime. Electronic modulation of the *I-V* relation rather than a mechanical process is required for picosecond resolution.

There are several ways to modulate the *I-V* curve on a picosecond time scale. One can generate a picosecond voltage pulse and couple it to the junction; or one can shine a laser pulse directly on the junction to bias with the voltage of the optical field. Either possibility appears plausible at first glance. I briefly consider each case below.

A1.2.1. Direct optical modulation

A 10 mW optical beam containing a train of 1 ps pulses at a 100 MHz repetition rate focused to a 1 μm spot size at the STM junction can produce (at most) 150 mV across a 10 Å gap. As discussed in Chapter 4, the STM junction *I-V* is given approximately by

$$I \propto V + bV^3,$$

where $b < 1$, and V is measured in Volts. For a DC bias of 1 Volt and a pump induced signal with a very large amplitude of 1 Volt, the signal from the V^3 term at the peak of the optical pulse constitutes on the order of 10% or less of the DC background signal. The measured signal is additionally reduced by the duty cycle of 10^4. Improving the time resolution decreases the signal strength. For every factor of 10 increase in time resolution, one gains a factor of $\sqrt{10}$ in voltage across the junction but loses a factor of 10 in duty cycle.

For the parameters considered above, the peak amplitude of the time resolved signal would be on the order of 0.1 pA. The signal above should be measurable in a stable STM junction. But 10 mW focused to a 1 μm spot produces gigantic thermal drifts. An estimate based on empirical results in the literature would predict vertical displacement of the tip relative to the sample on the order of 10 nm/ms when the laser is turned on[1]. A more reasonable spot size of 50 μm produces only on the order of a 5 mV bias across the junction, resulting in a few fA signal, which is probably impossible to measure over the shot noise and input noise of a preamplifier. In addition to thermal drifts, a large optical signal at the junction can generate spurious excitations on the sample which interfere with the processes of interest on the sample.

A1.2.2. Voltage pulse

The direct optical bias modulation described above, while potentially enabling high time resolution, is completely unviable in terms of signal to noise levels and junction stability. What about coupling an external voltage to the junction via some sort of optical switch? Such a scheme suffers the same signal amplitude problems as direct optical coupling, but one can ameliorate the effect by coupling larger and longer voltage pulses to the junction without significant adverse heating effects. Nunes and Freeman were able to get a measurable signal with signal to noise of about 10::1 by using large, (500 mV) long (100 ps) voltage pulses[2]. An increase in time resolution from 100 ps to 1 ps with the same laser repetition rate would result in a painful factor of 100 loss in signal strength due to the change in the duty cycle.

Appendix 1 References:

1. Nabil M. Amer, Andrew Skumanich, and Dean Ripple, "Photothermal modulation of the gap distance in scanning tunneling microscopy", *Appl. Phys. Lett.*, 49, p. 137 (1986). S. Grafstrom, J. Kowalski, R. Neumann, O. Probst, and M. Wortge, "Analysis and compensation of thermal effects in laser-assisited scanning tunneling microscopy", *J. Vac. Sci. Technol. B* 9, p. 568, (1991).

2. G. Nunes and M. R. Freeman, "Picosecond resolution in scanning tunneling microscopy", *Science*, 262, p. 1029 (1993).

Appendix 2
LabCentral PROGRAM INTERFACE

A2.1. Main screen and work area

Illustrated in Figure A2.1, the main screen of LabCentral appears when the program is started. One can then load stored graphs or acquire new data. In the screen of Figure A2.1, three graphs are loaded, each containing one or more traces drawn in several available styles. The graphs can be loaded from text files (containing ascii characters) or from binary files previously output from the program. The method by which the program reads text files can be customized to enable it to read graphical output from postscript generating programs like Mathematica. It is impossible to read such files into commercial data analysis software.

Once data is loaded, one can use the mouse to zoom in to different areas of a graph, measure the slope and distance between points, or overlay different traces. The style of each trace can be altered independently, and the

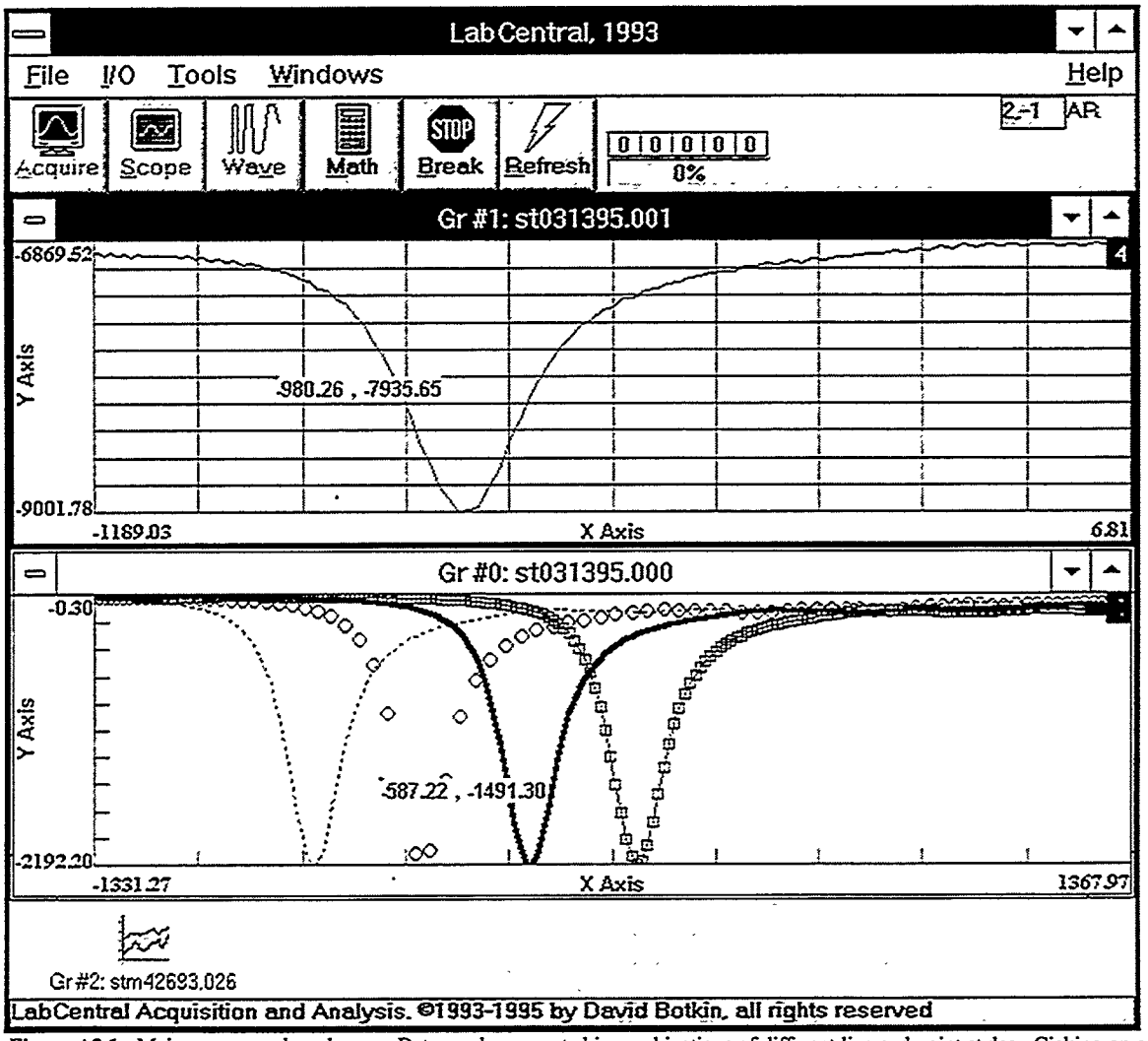


Figure A2.1. Main screen and work area. Data can be presented in combinations of different line and point styles. Cicking on a graph displays the coordinates at the clicking point. The mouse can also be dragged between two points to display slope and distance. The right moise button activates zooming. Graphs or curves can be overlaid by dragging one atop the other.

graphs can be printed exactly as they appear on the screen. One can also activate the calculator to manipulate data.

A2.2. Calculator

The calculator shown in Figure A2.2 functions as a normal calculator but transparently acts on arrays of data. Traces can be combined via standard arithmetic as well as via correlation and convolution. The interface is designed

to be visual; one simply clicks on the trace he wants to manipulate and clicks on the computation to be performed. The calculator can take an FFT, derivative, or integral of a trace, or can smooth or filter the data. Those who like to cheat can subtract arbitrary backgrounds or drag points in the data to new locations. Simple fits can also be performed, and one can generate arbitrary traces. The results of a calculation can be displayed by dragging the result of the calculation to an existing graph or to an empty space in the main window to generate a new graph.

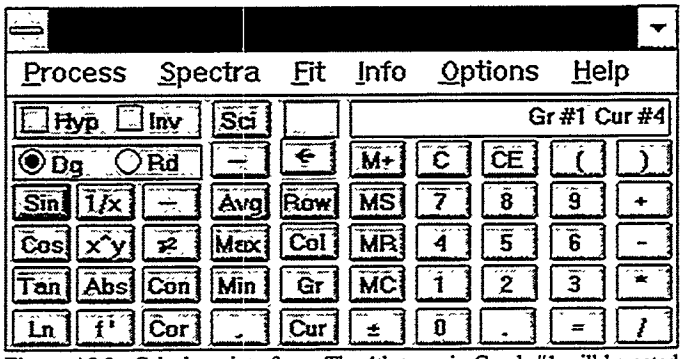


Figure A2.2. Calculator interface. The 4th trace in Graph #1 will be acted upon next.

While the basic tools to manipulate arrays of data exist in a variety of software packages, I have not encountered one which can be used as simply and intuitively as this one.

A2.3. Data acquisition

Figure A2.3 illustrates the setup screen for data acquisition. Scans can be performed with a National Instruments A/D card. The user can select the standard variety of options - points per scan, number of scans, and repetition rate. Sets of scans can be subtracted to remove baselines, and the signal measurement can be stopped if a discriminator signal leaves a specified range. These former feature is often used in optical absorption experiments. The latter is useful when there are noise spikes in the measurement.

Scans which involve operating or reading from instruments using the GPIB protocol also are supported. At the time of writing of this thesis, the GPIB routines were being upgraded and were not functioning. The planned upgrade will enable scans cutomized for any set of instruments - scanner and detectors- which use GPIB. The original routines only operated equipment used in our laboratory.

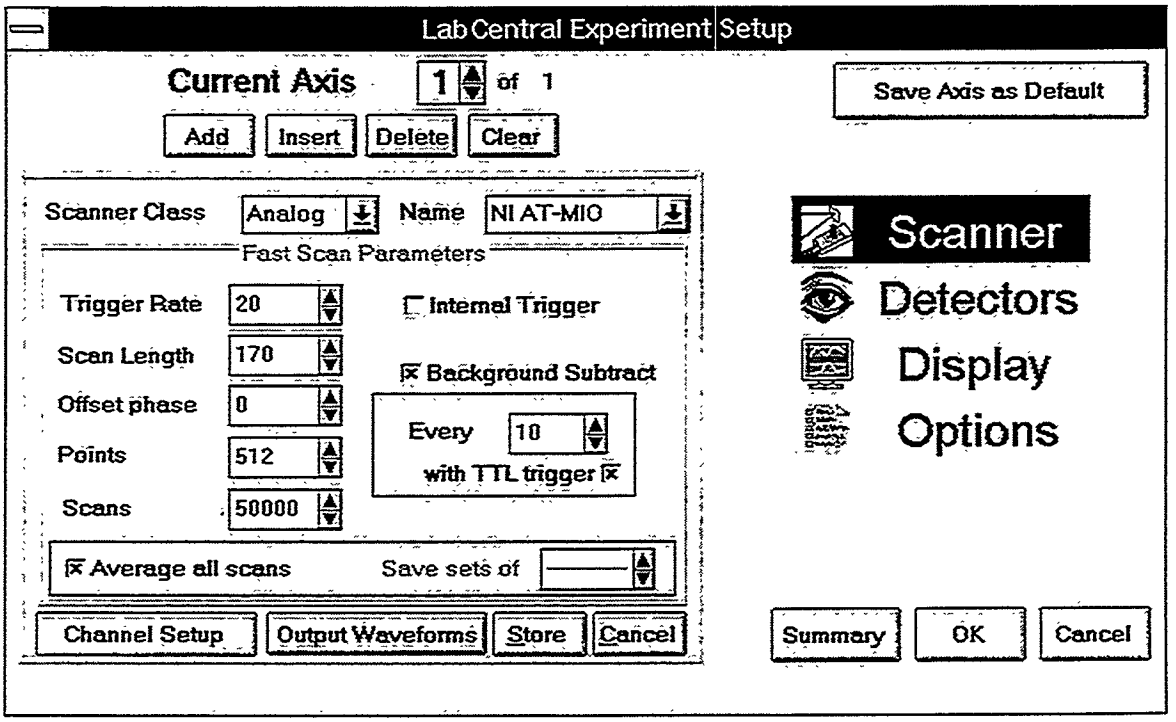


Figure A2.3. Setup screen for data acquisition. A/D and GPIB scans are supported.